Australian Government
Department of Defence
Defence Science and
Technology Organisation

A Review of Contributions by Australian Research Institutions into Speech Processing

Trevor Chi-Yuen Tao

Command, Control, Communications and Intelligence Division
Defence Science and Technology Organisation

DSTO-TN-0837

ABSTRACT

This report is a survey of contributions by various research institutions within Australia into several important applications of speech processing, such as speech and speaker recognition. The purpose of this report is to give a rough snapshot of where a number of individual research institutions stand. For each application, a number of research papers within Australia are discussed in detail. Although much of the above research is directed towards simple tasks there are a number of significant contributions from various Australian research institutions. Some systems, particularly those from QUT, have achieved state-of-the-art performance.

Published by

Command, Control, Communications and Intelligence Division
DSTO Defence Science and Technology Organisation
PO Box 1500
Edinburgh South Australia 5111 Australia

Telephone: (08) 8259 5555
Fax: (08) 8259 6567

AR-014-256
August 2008

APPROVED FOR PUBLIC RELEASE

A Review of Contributions by Australian Research Institutions into Speech Processing

Executive Summary

This report is a survey of how Australia has contributed to the research in speech processing. The customer is interested in which Australian research institutions have produced, or are likely to produce in the future, useful technologies for various speech processing applications. We have identified a number of functional components such as speaker verification, language identification, speech recognition etc. which can represent components of a larger system, although discussion of such a larger system is outside the scope of this report. For each of these functional components, we discuss the standard techniques used with emphasis on which techniques have been researched by Australian research institutions.

A large number of Australian institutions have done research on various speech processing applications and many interesting avenues of research have been investigated. However, a significant amount of their work focuses on irrelevant tasks (e.g. telephone banking) or problems that are too simple (e.g. 10-digit recognition) to be of use in high volume speech processing. Also, research experiments are often performed on a little-known corpus to validate their results, making it difficult to compare against other research. A notable exception is QUT, which is the only regular Australian participant in the well-known NIST evaluations, and their results are generally competitive. Another important contribution is the ANDOSL database, from four universities (excluding QUT). This database was used to promote research into database annotation and speaker diarisation.

This report is intended to highlight which institutions(s) are conducting research in areas of potential significance to the customer in providing support to their current and future high volume speech processing requirements.

Authors

Trevor Chi-Yuen Tao
Command, Control, Communications & Intelligence Division

Trevor Chi-Yuen Tao graduated from the University of Adelaide (Australia) in 2005 with a PhD in Applied Mathematics in 2005 and started employment at DSTO in August 2006. His current research involves speech processing.

Contents

Figures

Tables

1. Introduction

This review is a survey of speech-related technologies that have been undertaken by various Australian institutions. It is intended to help a client decide which research institutions in Australia can be considered for funding opportunities. In broad terms, speech processing is the study of speech signals and the algorithms applied to them. This report will discuss various systems that perform a specific function, such as speaker verification, language identification, keyword spotting etc. It is anticipated that a number of such systems will form part of a larger end-to-end system, but this is outside the scope of this report. A system will generally contain a number of operational components such as feature extraction, phoneme modelling, model training etc. This report will describe in detail various operational methodologies used by Australian institutions in a number of selected sub-problems. The subsequent chapters cover, in order, Speaker Verification (SV) and Identification (SID), Language Identification (LID), Speech Recognition, Keyword Spotting (KWS), Accent Identification (AID), phoneme recognition, speech segmentation and other lesser-known topics. It will be noted that some components are omitted despite being popular fields of research. For instance, speech coding, speech enhancement or natural language understanding will not be discussed.

2. Speaker Recognition

2.1 Overview of Speaker Recognition

In speaker recognition, two important sub-problems will be considered, namely speaker verification (SV) and speaker identification (SID). The aim of SV is to determine if a particular speaker was speaking or not given a speech segment. A SV system generally has two phases, training (offline) and testing (online), and consists of the following components: parameterization, modelling and scoring (see Figure 1). Parameterization occurs in both the training and testing phases, while modelling and scoring occur only in the training phase and testing phase respectively.

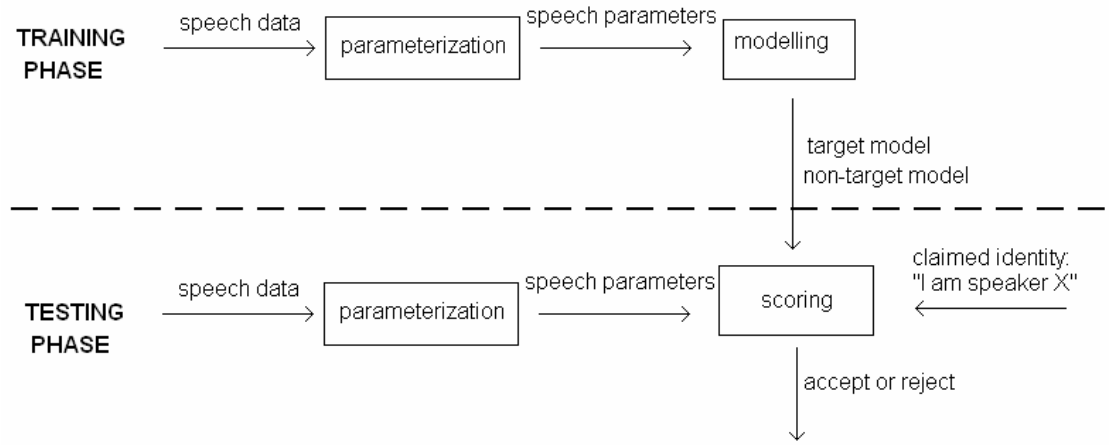

Figure 1: Modular representation of speaker verification system

Parameterization means analysing the speech signal and converting the raw data into features, such as Mel Frequency Cepstrum Coefficients (MFCC) or Linear Prediction Cepstrum Coefficients (LPCC) etc. The parameterization process is used both in training speaker models and testing (see below). Once the parameters are obtained, speaker models can be trained using a sufficiently large quantity of data. In the testing phase, speech data is converted into features in the same way as in the training phase. One is also given a claimed identity, from which one can obtain the target model from the training phase. The target model represents the hypothesis that a given audio segment was spoken by a particular speaker. A Universal Background Model (UBM) is obtained from pooled speech by several speakers. This model represents the alternative hypothesis that an audio file was not spoken by a particular speaker. Using a likelihood-ratio test a "score" is obtained, which indicates how confident one is about whether the speaker should be accepted. The higher the score is, the stronger the evidence is in favour of accepting the speaker. Note that an individual score is not to be confused with the result of an evaluation (such as specified by NIST) where an algorithm is tested against some data and a result is given, e.g. "given task XXX and corpus

YYY, algorithm ZZZ makes correct decisions 80% of the time". The scoring stage determines whether the claimed identity should be accepted or rejected. Typically this is done by comparing the score against a pre-determined threshold. If the score exceeds the threshold, the speaker is accepted, or else it is rejected. This threshold can be viewed as a parameter, controlling the trade-off between false alarms (where an impostor is accepted) and misses (where the speaker is incorrectly rejected). Increasing the threshold will result in a lower false alarm rate at the cost of a higher miss rate, while decreasing the threshold yields the opposite result. Thus one can construct a Detection Error Trade-off (DET) curve that shows how the false alarm rate and miss rate change as a function of the threshold. The score is often normalized (before comparing to a threshold) to neutralize certain effects such as environmental noise and mismatched training/testing conditions (e.g. carbon versus electret handsets), since it is well-known that such effects can greatly reduce performance.

A number of evaluation metrics are commonly used to assess SV systems. The basic idea is to summarise the DET curve in a single value by choosing an "operating point" (corresponding to a value of the threshold) which is optimal in some sense. The cost function assigns a real number to both misses (C_{miss}) and false alarms (C_{fa}). The total cost is given by $C = C_{fa} P_{fa} + C_{miss} P_{miss}$. Note that the impostor rate must be assumed to be known a priori to enable the calculation of false alarm and miss probabilities. The desired operating point is where the cost function attains the minimum value C_{opt}. This value summarises the performance of the system, with lower values of C_{opt} indicating better performance. Another measure commonly used is the EER, where the probabilities of false alarm and miss rate are equal. However, this is less popular than the cost function, since the EER rarely corresponds to a realistic operating point (Bimbot et al., 2003). Other measures exist, but it is outside the scope of this report to discuss these.

In SID, the problem is: given a set of speakers (generally a finite set, known a priori) and a speech signal, determine the most likely speaker. The procedure for SID is somewhat similar to SV except (i) all speaker models obtained in the training phase are required for scoring, (ii) no claimed identity is required and (iii) in the scoring stage one chooses the speaker which corresponds to the highest-scoring model (see Figure 2). This is simpler in the sense that one does not need to select an arbitrary threshold or perform normalization techniques such as T-norm (as in SV). This implies that scoring is relatively trivial, and the speaker with the highest score is chosen.

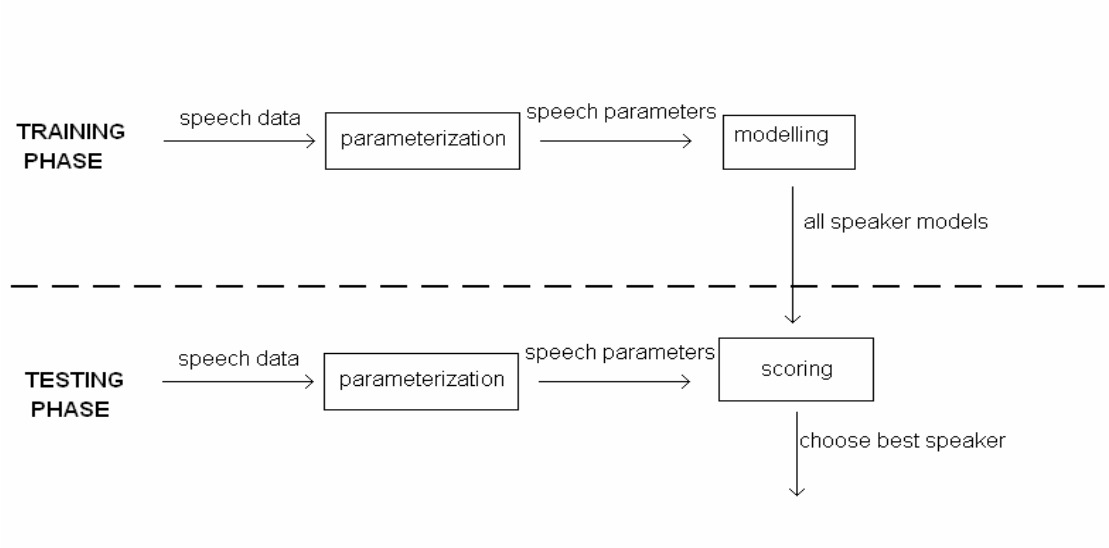

Figure 2: Modular representation of speaker identification system

Section 2.2 describes in detail a number of ideas by Australian research institutions for the parameterization stage, such as Sliding Window Analysis (Pelecanos & Sridharan, 2001b) and Feature Warping (Pelecanos & Sridharan, 2001a). Section 2.3 describes some models proposed for the modelling stage. Section 2.4 offers only a brief discussion on scoring, since this has not been heavily researched by Australian research institutions.

2.2 Parameterization

Parameterization essentially consists of two stages: feature extraction and filtering/normalization. Feature extraction involves preprocessing the signal and selecting a parameterization scheme to generate various features of interest. Given a speech signal, one can apply standard techniques such as FFT, multiplication by filter-bank etc, to generate spectral vectors. These can be further processed to generate, for instance, the well-known Mel Frequency Cepstrum Coefficients (MFCC) or Linear Predictive Cepstral Coefficients (LPCC). The filtering/normalization stage is needed to remove noise and compensate for "mismatched conditions" such as different duration of syllables. A well-known method for filtering is Cepstral Mean Subtraction (CMS) (Atal, 1974) where the mean vector is subtracted from each cepstral vector. A number of ideas from various Australian institutions will be discussed below.

2.2.1 Sliding Window Analysis

In (Pelecanos & Sridharan, 2001b) from QUT, Pelecanos and Sridharan proposed to filter away noise using "sliding window analysis". That is, a box-car filter is applied to the cepstral features as a function of time and the output of this filter is subtracted from the raw cepstral features. This generalizes the well-known CMS filter in the sense that the mean is subtracted from a small "sliding" window instead of the entire speech segment. The great advantage of sliding window analysis is that unlike CMS, window analysis is suitable for handling multiple

channels, e.g. when both sides of a telephone conversation are recorded. Although other well-known filters have been designed to handle multiple channel effects, such as RASTA (Hermansky & Morgan, 1994) and LDA-FIR modulation spectrum analysis (Avendano, Van Vuuren, & Hermansky, 1996), they have some drawbacks**.** RASTA only shows limited improvement over CMS. Moreover, there is the issue of settling time of the Infinite Impulse Response (IIR) filter at the start of a speech segment, which can degrade performance for short test segments (of the order of 3 seconds). The data-driven approach of LDA-FIR requires the use of phonemically transcribed speech segments that are compatible with the speech in the target application.

Sliding window analysis avoids all these difficulties. Pelecanos and Sridharan showed that a window of 300-500 frames is optimal for the two-speaker detection task in the NIST 2000 speaker recognition evaluation (A. Martin & Przybocki, 2000).

2.2.2 Feature Warping

Another solution to linear channel effects/additive noise is feature warping (Pelecanos & Sridharan, 2001a). The use of this technique was the key technology in QUT gaining first place in the 2001 NIST world-wide speaker recognition evaluation in the Single Speaker Detection Task (Basic) and Single Speaker Detection Task (Cellular Data) categories. Feature warping attempts to enhance the robustness of each cepstral feature distribution by mapping it to a target distribution, such as a standard normal distribution. At any particular time, each cepstral value is "ranked" according to how many other values in a sliding window exceed it. The corresponding percentile in the target distribution becomes the warped cepstral feature value. The effect of feature warping is to emphasize the *relative* instead of absolute values of the feature vectors. A noteworthy aspect of feature warping is that it can be cascaded with other schemes such as RASTA (Hermansky & Morgan, 1994). Feature warping is related to the concept of histogram equalization, a well-known technique for image processing, and has been widely adopted in the speech processing community. For instance, Choi, from ATP Research Laboratory, NICTA, (Sydney) used it for robust front-end processing for speech recognition (Choi, 2006).

2.2.3 The Use of F-patterns for Diphthongs

Another useful set of features can be obtained from the F-patterns (fundamental frequency F0, first second and third formants F1 F2 F3 etc). Rose (P. Rose, 2006), from ANU, postulated that diphthongs carry more useful information than monophthongs. Rose explored two questions: (i) given two speech samples, to what extent diphthongs can be used to discriminate whether they are from the same or different speakers and (ii) which parameters are appropriate for which diphthongs. For the diphthong /ai/, Rose showed that including the F2 + F3 formants, omitting the F1 formant and normalizing the duration gave the best EER of 22% on the Bernard Corpus (Bernard, 1967). Other different diphthongs such as /ɛi/ have yielded similar results. The diphthong /ai/ has useful properties for forensic SID (P. Rose, Kinoshita, & Alderman, 2006): the three formants are relatively easy to measure and often occur in phone conversations (e.g. 'Hi', 'Bye'). Rose also showed that combining five different diphthongs can improve the EER to 10% for the Bernard Corpus. Although the power of formants has been

researched in the context of monophthongs (Alderman, 2005; P. Rose, Osanai, & Kinoshita, 2003), diphthongs remain relatively unexplored, especially in the context of forensic applications, so this could be a fruitful area for future research. The Bernard corpus is relatively simple, as Bernard recorded only the occurrences of diphthongs in /h_d/ words. Hence this is only useful in text-dependent SID, where the given text is known. A more powerful system would require at least a phoneme-recognition stage to handle unknown text.

2.2.4 Discriminative feature extraction

In speaker recognition, as well as other areas of speech processing in general, a standard assumption is that the feature extraction algorithm is fixed while a classifier is adapted during training. In (Nealand, Bradley, & Lech, 2002) from RMIT University, Melbourne, Nealand et al. proposed that Discriminative Feature Extraction (DFE) could be used to train feature extraction parameters in conjunction with the classifier. The feature extraction employed a filter-bank-based extraction algorithm. The filter-bank was emulated by a weighted summation of power spectral components, where the individual weights could be trained.[1] The term "discriminative feature extraction" is derived from the fact that DFE attempts to maximize discrimination between classes, rather than fitting classifier models to training data. Nealand et al. showed that their DFE algorithm consistently scored a higher recognition rate than a conventional algorithm using a fixed filter-bank.

2.2.5 Higher level features

In the previous discussion only low level features have been considered. However, it is known that high-level features such as linguistic context, prosodic cues etc, also carry useful information and recently more researchers have investigated these sources of information for speaker recognition. In (Baker & Sridharan, 2006), from QUT, Baker and Sridharan used a multi-lingual framework where phones can be categorized into one of four broad classes, namely (i) vowels/diphthongs, (ii) nasals/liquids/glides, (iii) fricatives and (iv) stops/pauses. A pseudo-syllable was assumed to consist of three phones. Thus pseudo-syllables could be classified as one of only 64 possible combinations of three broad phone classes. This represents a reasonable trade-off between the ability to retain useful information in the phone classification set and to allow for sufficient training data for each pseudo-syllable. By modelling these broad syllabic events, comparable performance to a standard system was obtained on the NIST 2003 speaker recognition corpus.

However, attempts to incorporate high-level features such as prosody to complement the standard acoustic features have met with limited success, both within Australia and worldwide. In Luengo et al. (Luengo et al., 2006) a traditional MFCC-based SV system was combined with a prosody-based system and tested on the AHUMADA database (Ortega-Garcia et al., 1998). The traditional and prosody-based systems obtained an EER of 3.85% and 23.93% respectively, hence the prosody-based system is much inferior. When the two systems are combined the EER drops to 3.84%, which is a negligible improvement over the MFCC-based system alone. One of the major issues with higher-level features is the necessity to

[1] It is assumed the filter has a Gaussian profile to reduce the number of trainable parameters.

mark-up large volumes of training data with higher-level feature information such as prosodic features.

2.3 Modelling

2.3.1 Gaussian Mixture Models

Text-independent SID and SV is a strong focal point of research in Queensland University of Technology (QUT) as evidenced by regular participation in NIST evaluations and numerous publications in conference proceedings, particularly ICASSP. QUT has been particularly successful with the use of Gaussian Mixture Models (GMM). The GMM is relatively simple and well-understood, yet other more complex models such as HMMs and Neural Networks (NN) have failed to demonstrate any consistent advantage over GMM. For this reason GMM is considered one of the most successful models in SID and SV (Bimbot et al., 2003). A GMM is a likelihood function that specifies the probability density for a feature vector (e.g. MFCC coefficients) to be a linear combination of Gaussians with weights summing to unity. The GMM parameters are iteratively updated using the expectation maximization algorithm (Dempster, Laird, & Rubin, 1977) during model training. Typically, the GMM representing a speaker is not created "from scratch" but is adapted from a Universal Background Model (UBM) representing all speakers (Reynolds, Quatieri, & Dunn, 2000).

2.3.2 Trajectory Modelling

The HMM model is based on the assumption that speech is "sustained" in one state before suddenly jumping to the next state. But this assumption is not realistic in continuous speech since sustained sounds are short or omitted. In (Tey, Jong, & Togneri, 1996), from UWA, a speech signal was treated as a "moving point" in N-dimensional space and segmented, using a transient trajectory model for each transition from one sound to another. The trajectory of a speech signal was viewed via a Feature VIEWing system, `fview`, a software package developed within CIIPS (Centre for Intelligent Information Processing Systems), in the School of Electrical, Electronic and Computer Engineering of UWA. With the help of this package, a speech signal was manually segmented. For each segment, each cepstral coefficient (as a function of time) was modelled as a low-order polynomial. Another idea proposed by the same authors was that of "rate-independent" parameterization. Effectively, the time dimension was "warped" so that the speech signal moves in N-dimensional space at a constant rate (with respect to the standard Euclidean distance metric) in order to normalize differences between different speaker rates. The experimental results were disappointing: it was found that the trajectory model was comparable to HMM for speech recognition but inferior for speaker recognition. Moreover, the model was only tested on a simple set of sounds, namely the English letters "b,c,d,e,g,p,t". The experimental results favoured the normal "unwarped" rate-dependent parameterization.

Despite the poor experimental results this paper is noteworthy since UWA has placed `fview` in the public domain. It has been used for research into speech recognition and speaker recognition, but its main purpose was to promote research into front-end processing as an individual component, not within a system. Tey et al. (Tey, Jong, & Togneri, 1996) noted that

while it is common to publish results for various systems (LID, speech segmentation etc), of which front-end processing is only one component, little effort has been devoted to considering front-end processing in itself. Thus it is difficult to determine if a system "does badly" because of weaknesses in the front-end processing component, or some other component in the system. Unfortunately, outside of UWA, there is no significant research dedicated to `fview`.

2.3.3 Neural Networks

In (Price, Willmore, Roberts, & Zyga, 2000), from DSTO, Price et al. compared a NN model against a conventional GMM. Their work is novel in that genetic algorithms are used for optimising the network. An individual NN was created to model each speaker. Each NN took cepstral coefficients as the input feature vector and returned a binary output of 10 or 01 to represent the speaker and background respectively.[2] The training data consisted of 2 minutes of speech from 2 different handsets for 21 different male speakers from the 1996 NIST speaker evaluation workshop. The test data utterance length was nominally 30 seconds. Price et al. trained and tested both matched and mis-matched microphone conditions. It was found that GMM outperformed NN in terms of equal error rate under both matched and mis-matched conditions. Given a miss rate of 0.5% GMM also outperformed NN (9% vs 14% probability of false alarm) under matched conditions. But under mis-matched conditions the result is reversed, (98% vs 64% probability of false alarm).

2.3.4 Other Models

Other methods for solving the classification problem include HMM, and SVM. HMMs attempt to incorporate temporal information by using transitions between states to model how a signal evolves with respect to time. SVMs attempt to separate speaker and impostor models without the assumption of a linear boundary. Attention has also been paid to combining GMM with SVM. For a detailed discussion the reader is referred to (Bimbot et al., 2003) and references therein.

In (Baker & Sridharan, 2006) (see section 2.2.5) it was assumed that a syllable consisted of three phones. The HMM is appropriate since one can easily concatenate HMM models for individual phone classes to form a model for the entire syllable. Baker and Sridharan used a 7-state left-to-right topology, since each phone was modelled as a 3-state left-to-right topology and the entry and exit states of the middle phone overlapped with those of the start and end phones. The HMM was tested against a previously proposed GMM system (Baker, Vogt, & Sridharan, 2005) and a "baseline" GMM-UBM system (Reynolds, 1997) on the NIST 2004 speaker recognition evaluation corpus. It was shown that HMM outperformed the GMM and GMM-UBM systems. This result is potentially significant. However it would be computationally slow due to the requirement for phoneme recognition to be performed.

[2] Although it is possible to return only one bit instead of two, it is generally accepted that two bits yield better performance.

2.4 Scoring

The scoring component determines whether a speaker should be accepted or rejected depending on the speaker's score. There are two distributions representing the target and impostor models. It is assumed that both models are approximated by normal distributions. Usually the output score is compared to a threshold. If the score is higher than the threshold, then the speaker is accepted, otherwise it is rejected. Score normalization is often used and is based on the following idea: instead of directly comparing the score θ to a threshold, it is better to compare $(\theta - \mu)/\sigma$ with the threshold, where μ, σ are the mean and standard deviation of the target model respectively. The initial study of score normalization techniques is largely due to Li and Porter (Li & Porter, 1988). There are a large number of variations on the theme of score normalization and the reader is referred to (Bimbot et al., 2003) for detailed discussion. QUT has used score normalization techniques from other researchers outside Australia. For instance, Mason et al. (Mason, Vogt, Baker, & Sridharan, 2004) have used handset normalization and test segment normalization (Auckenthaler, Carey, & Lloyd-Thomas, 2000). However the author of this report is unaware of any serious contributions by Australian research institutions to the use of score normalization *per se*.

3. Language Identification

3.1 Overview of Language Identification

The aim of Language Identification (LID) is to determine a language given a speech segment. A typical LID system consists of the following stages: parameterization (feature extraction), modelling and scoring (see Figure 3). Hence it is similar to speaker identification, but with "speaker" replaced by "language". Parameterization and modelling will be discussed in some detail.

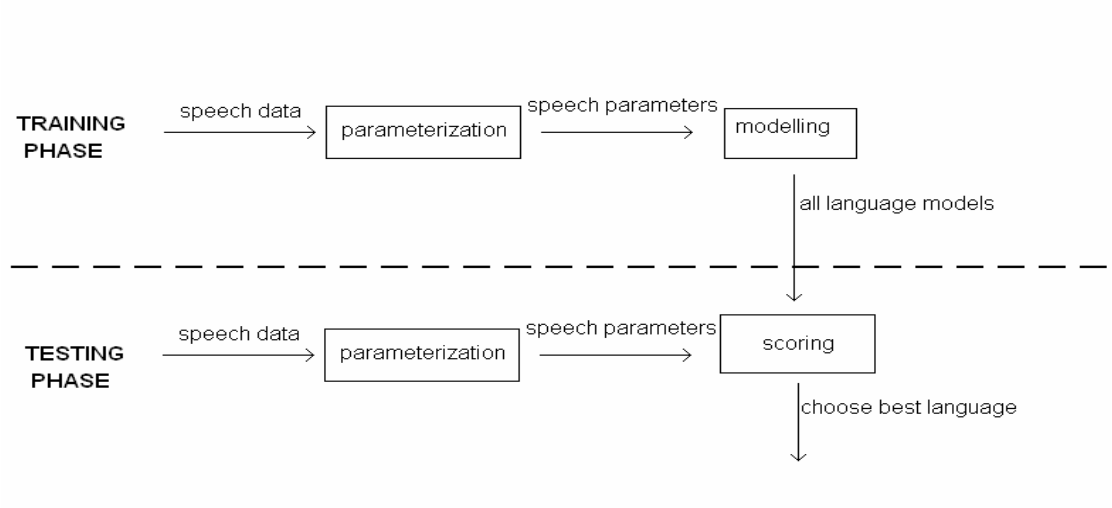

Figure 3: Modular representation of language identification system

3.2 Parameterization

3.2.1 Mel Frequency Cepstral Coefficients (MFCC) versus Linear Predictive Cepstrum Coefficients (LPCC)

Mel Frequency Cepstral Coefficients (MFCC) and Linear Prediction Cepstrum Coefficients (LPCC) are two of the more common parameterization schemes. The MFCC are obtained by computing log output amplitudes of non-linearly spaced filters and taking a discrete cosine transform. LPCC attempts to model the vocal tract by an all-pole filter. The linear prediction coefficients (LPC) are the coefficients of this filter and the LPCC are the same data in the cepstrum domain. The name LPCC derives from the fact that future values of the signal are modelled as a linear function of previous samples. In (Wong & Sridharan, 2001), from QUT, Wong and Sridharan compared LPCC with MFCC. They showed that LPCC consistently outperformed MFCC in all tests using the 10-language version of OGI-TS. Also, using delta coefficients (time derivatives of the features coefficients) resulted in enhanced performance. Wong and Sridharan used the GMM for the modelling component, as described above in the chapter on speaker identification. In the experiments performed, the optimum accuracy was only 60.0% achieved with the use of 12 LPCC and corresponding delta and acceleration

coefficients. Unlike speaker identification, the use of GMM is less favourable compared to other methods, such as phoneme-based modelling (Zissman & Singer, 1994). This will be discussed in more detail in section 3.3.

3.2.2 Prosody

In Martin et al. (T. Martin, Baker, Wong, & Sridharan, 2006; T. Martin, Wong, Baker, & Mason, 2004) phone-triplets were used as crude approximates for a syllable-length sub-word segmental unit. This pseudo-syllabic length framework was subsequently used to examine the contributions made by acoustic, phonotactic[3] and prosodic information sources to gain insight into how these information sources contribute to overall LID performance. Importantly, this work was examined under current NIST LID evaluation protocols, in conjunction with typical baseline LID techniques such as the GMM/UBM and PPRLM approaches.

A series of experimental comparisons were conducted, examining the utility of segmental units in modelling short term acoustic features. This included comparisons between language specific GMM, language specific GMM for each segmental unit, and finally language specific HMM for each segment. This examination was undertaken in an attempt to better model the temporal evolution of acoustic features. In a second tier of experiments, the contribution of both broad and fine class phonotactic information, when considered over an extended time frame, was contrasted with an implementation of the currently popular parallel phone recognition language modelling (PPRLM) technique. Results indicated that this information could be used to complement existing PPRLM systems to obtain improved performance. The pseudo-syllabic framework was also used to model prosodic dynamics and compared to an implemented version of a recently published system, achieving comparable levels of performance.

Further studies examining the use of prosody for LID were also conducted in Bo et al., (Bo, Ambikairajah, & Fang, 2006) from UNSW and NICTA. In this study, prosodic information was combined with cepstral features such as MFCC and PLP. In contrast to most studies which utilise 12 MFCC or 9 PLP coefficients, Bo and Chen also investigated the effect of altering the number of coefficients. They reported an optimum performance of 87.1% on the 10 language recognition task in the 1992 OGI corpus, obtained when the number of MFCC components was reduced to seven.

3.2.3 Acoustic Systems

Recently, acoustic LID has received renewed interest due to the NIST 2003 evaluation task (Matejka, Cernocky, & Sigmund, 2004) in which acoustic-based systems outperformed traditional phonetic systems. In (Allen, Ambikairajah, & Epps, 2006), again from UNSW and NICTA, Allen et al. combined information from both magnitude and phase information in the signal. The phase information was obtained via the Modified Group Delay Function (MGDF) (Hegde & Murthy, 2005). The use of the Shifted Delta Cepstrum (SDC) (Torres-Carrasquillo et al., 2002) further improved performance. Instead of calculating an approximate first derivative

[3] Roughly speaking, phonotactic information refers to how often different combinations of phonemes can appear in a language.

(Furui, 2000), a vector of first order delta coefficients was obtained by computing differences across multiple frames of speech.

Curiously, the MGDF had also been used by Griffith University in the context of speech recognition (Donglai & Paliwal, 2004), although that paper only focussed on the relatively simple task of recognising connected digits. Phase information is relatively unexplored in the literature yet has interesting properties. For instance, speech signals can be reconstructed from the magnitude spectrum only if the phase information can be estimated accurately.

3.3 Modelling

3.3.1 Hybrid Systems

Although the use of GMM in LID compares unfavourably with other methods, GMM can be incorporated into "hybrid systems" using other more successful techniques. In (Wong & Sridharan, 2002a), from QUT, Wong and Sridharan showed that the use of Voice-Tract Length Normalization (VTLN) (Wong & Sridharan, 2002b) and Parallel Phoneme Recognition and Language Modelling (PPRLM)[4] (Zissman, 1995; Zissman & Singer, 1994) can significantly reduce error rates of GMM. Since PPRLM is one of the more successful approaches, it will be discussed in further detail in section 3.3.2.

The UBM technique, successfully employed in speaker verification, was employed in LID to reduce the computational costs in both training and testing of the GMM model. VTLN, which has proved very successful in speech recognition (Bacchiani, 2001; Wegmann, McAllaster, Orloff, & Peskin, 1996), attempts to normalize out the inter-speaker differences due to variable vocal tract length. Wong and Sridharan incorporated VTLN into their GMM system by effectively attempting to adapt the GMM model during the training phase and estimate the tract length[5] of each speaker simultaneously (Wong & Sridharan, 2002b). Hybrid systems are generally implemented as follows: given a language, each system outputs a score. The resulting scores by all models are then "combined", usually via a weighted average (Wong & Sridharan, 2002a). The scores for all languages are used in the final decision of identifying a language. Wong and Sridharan attempted to fuse the GMM with the phoneme-based approach PPRLM**,** the latter having proved very successful in LID. Given output scores from GMM and PPRLM, the final output was a linearly weighted average of the two (although the weights heavily favour GMM). Wong and Sridharan showed that the fused system outscores both GMM and PPRLM separately. However, it should be pointed out that their experiments also confirmed that GMM by itself is inferior to PPRLM, even when the UBM and VTLN techniques are applied to the former. These results suggest that the static acoustic features identified by the GMM somehow "complement" the phonemic information captured by PPRLM. In (T. Martin, Baker, Wong, & Sridharan, 2006) (section 3.2.2) a similar idea was applied: the proposed syllable-based system was fused with acoustic HMM, prosodic HMM and PPRLM systems. However, instead of a simple linear weighting, a MLP neural network

[4] The acronym PRLMP also appears in the literature.
[5] Strictly speaking the authors merely estimate an "abstract" rather than the actual "physical" tract length of the speaker, but this distinction will be ignored.

was used. Martin et al. showed that the combination of these systems results in a significant improvement over the use of PPRLM alone.

Since it is outside the scope of this report to detail all of the above modelling techniques, only a brief discussion on PPRLM is provided. The interested reader is referred to (Muthusamy, Barnard, & Cole, 1994) and (Zissman, 1996) and references therein.

3.3.2 Phoneme Recognition and Language Modelling in Parallel

The concept of phonemes is fundamental to many (but not all) speech processing tasks. Phonemes are used in LID, Keyword Spotting (KWS) and Accent Identification (AID). One of the main advantages of phonemes is their small number compared to the set of words in the vocabulary. For instance English has only around 40 distinct phonemes but the number of words is higher by several orders of magnitude.[6] Moreover, as all words can be expressed as combinations of terms in the distinct phoneme set, phoneme models need no retraining when words are added to the vocabulary. Given models for all phonemes, any sentence can be represented by concatenating the word models, which can be obtained by concatenating the phoneme models. Although units other than phonemes have been proposed in the literature (Lee, Soong, & Paliwal, 1996), it is outside the scope of this report to discuss them in detail.

One of the most successful approaches to the modelling component of an LID system is PPRLM. Given the acoustic representation, phoneme models are constructed for multiple languages in parallel. A phoneme model is typically represented as a HMM with left-to-right topology to account for the temporal aspects of the phoneme. Phonemic labelling of data is often performed manually, despite being tedious and error prone, since accurate labelling is considered critical for system performance. For many databases, phonemic transcriptions are available for only some of the languages. Given an acoustic model, one can build a stochastic grammar for any language (not necessarily the same as that of the acoustic model). The likelihood of any combination of language model and acoustic model can be calculated. The likelihood of a language model is obtained by summing the likelihood of all possible combinations of the language model with any acoustic model. Scoring is usually done by choosing the language with highest likelihood. This process is depicted in Figure 4.

[6] Although several estimates have been given for the number of words in the English language, they will not be quoted since there is no single sensible criterion for counting words.

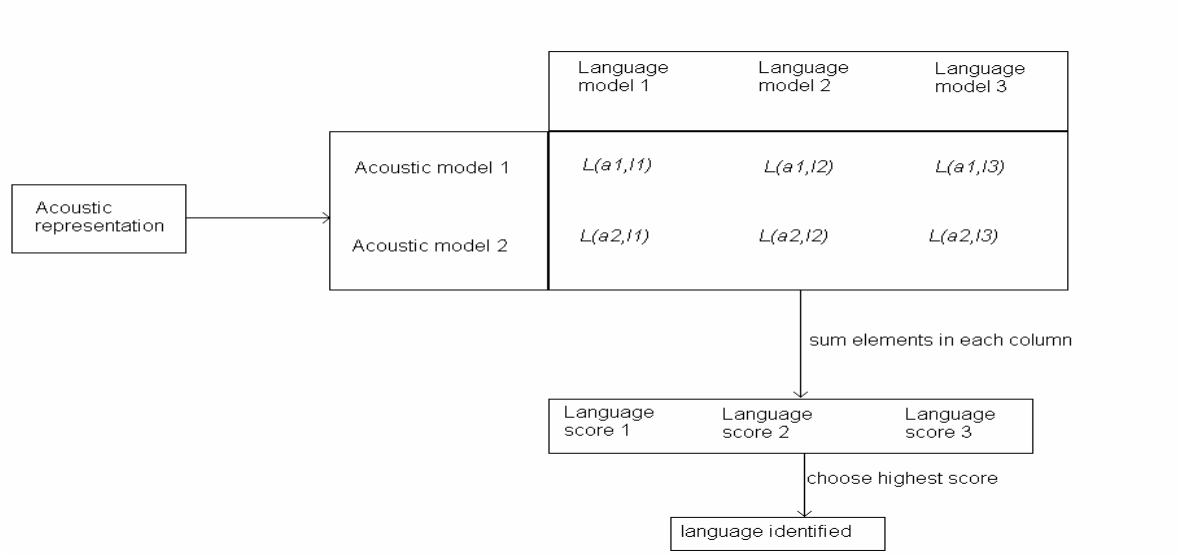

Figure 4: Data flow for PPRLM method

Given that the objective of PPRLM systems is to capture as accurately as possible the phonotactics which characterise a language, it is assumed that the minimisation of Phone Error Rate (PER) is a precursor to achieving this effectively. In (T. Martin, Wong, & Sridharan, 2006), Martin et al. investigated the relevance of PER as a metric for determining eventual LID performance. In contrast to previously reported techniques using PPRLM, this study made use of the CallHome corpus to produce the acoustic models, rather than OGI, based on the premise it provides a better representation for the style of discourse and channel conditions encountered in the Conversational Telephone Speech (CTS), which is now the focus of current NIST LID evaluations. Using the CallHome corpus, significantly improved results were obtained, with an average improvement of approximately 6% absolute across the 30, 10 and 3 seconds tasks for the NIST 1996 and 2003 evaluations. An examination was also conducted into the impact of tuning the individual front-end recognisers, on both the resultant PER of other languages and against the resultant LID performance. The work conducted established a number of limitations in the correspondence between PER and LID so a new technique based on pronunciation modelling techniques was trialled for forecasting the change in LID performance when the phone recogniser front-end was modified. The essential idea was to assign a smaller cost for mistaking two similar phones (e.g. /b/ and /p/) than mistaking two different phones (e.g. /b/ and /s/). It was shown that this measure gives stronger correlation to LID performance than PER.

4. Speech Recognition

4.1 Overview of Speech Recognition

Automatic speech recognition is the problem of decoding a speech stream into a sequence of words. Speech recognition applications range from extremely simple tasks (ten-digit recognition) to more complex tasks (e.g. medium or large size vocabulary voice dictation). Speech recognition is a well-researched area largely thanks to the advances in signal processing theory, algorithms, software and hardware. However, current technology is still significantly behind human performance under "real world" conditions. For instance, recognition accuracy can significantly degrade when the testing and training sets have different characteristics such as additive white noise or different speaking styles. Many Australian research institutions have done research into speech recognition but most of it is relevant for only small-vocabulary applications. This section will only cover medium-large vocabulary continuous speech recognition as this is particularly relevant for defence applications. It is obvious that the use of a large vocabulary and continuous speech will make speech recognition much harder. A large vocabulary may lead to infeasible constraints in memory or time. In other words, a system that works on a small vocabulary may not scale up to a larger vocabulary. Continuous speech implies that individual words may be pronounced differently depending on the context of neighbouring words, or a person may stutter or retract 'false starts' etc.

Given a speech signal, Large Vocabulary Continuous Speech Recognition (LVCSR) attempts to decode it into a sentence (or a number of sentences). Note that in some applications, further processing may be needed after obtaining sentences (e.g. speech understanding) but only the speech recognition problem will be considered. A speech recognition system consists of a parameterization module and a pattern-matching module. The parameterization module obtains a stream of features, just as in speaker or language identification. The pattern searching model takes these input features and decodes them into words/sentences, using word and sentence matching. Note that the word and sentence matching submodules are "coupled" because both modules give some measure of how likely a feature stream represents a particular sentence: the word match module determines how well each individual word matches the feature stream, and the sentence match module determines how well the individual words fit into a sentence. One possible interpretation is that the word module does not yield a hard choice of any particular word, but a set of probabilities for multiple words. Thus, for instance, the word "*The*" could be determined as the most likely word to begin a sentence, but after the rest of the signal is decoded the sentence module may indicate the sentence most likely starts with "*They*" instead. The word and sentence matching modules output a probability for any word sequence. Finally, a decoding module is required to calculate the word sequence with highest probability (see Figure 5).

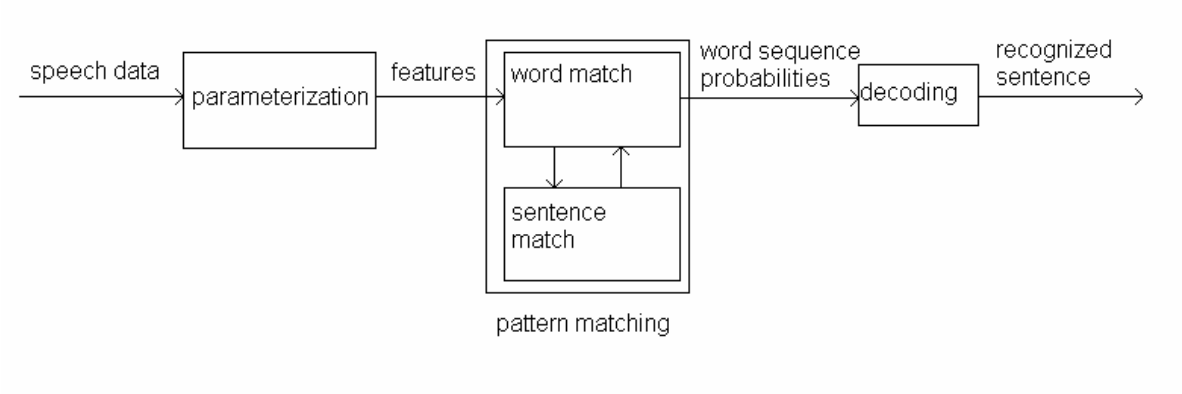

Figure 5 Modular representation of speech recognition system

4.2 Parameterisation

The parameterization stage in speech recognition has much similarity with the features used in Speaker Recognition. Fourier analysis is one of the most widely used tools for deriving speech features. Cepstral features are often used, along with their first and second derivatives. Mel- or Bark-scale spectral features are also common. For a detailed discussion the reader is referred to (Lee, Soong, & Paliwal, 1996). Feature extraction techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) have been used for the problem of audio-visual speech recognition (Dean, Lucey, & Sridharan, 2005; Dean, Lucey, Sridharan, & Wark, 2005), but for the remainder of this report it is assumed that only the audio signal is given.

4.2.1 Subband Spectral Centroid Histograms

In (Gajic & Paliwal, 2006; Paliwal, 1998), from Griffith University, Paliwal and Gajic used Subband Spectral Centroid (SSC) histograms The main motivation was to provide robustness against white noise. The SSC is related to spectral peak positions, but easier to compute. The SSC is therefore robust to additive noise, provided the noise spectrum is relatively flat and assumed independent of the signal. Paliwal and Gajic showed that SSC is comparable with MFCC in the presence of additive white noise, and comparable in noise-free conditions. Although their experiments were performed only on a small to medium size vocabulary, it is not inconceivable that the use of SSC can be applied to LVCSR.

4.3 Word and sentence matching

The Hidden Markov Model is the most commonly used model in continuous speech recognition systems. The HMM is used to represent phonemes, which are then concatenated sequentially to represent words and/or sentences. One main advantage of the HMM approach is its ability to decode a temporal sequence without need for manual segmentation, which is very tedious. On the other hand it is difficult to justify the ad-hoc choice of a particular model. This amounts to the assumption that the observed speech is produced by a particular underlying distribution and it is only necessary to estimate model parameters of

such a distribution. This is considered to be one of the main drawbacks of HMM (Bahl, Brown, de Souza, & Mercer, 1993; L. R. Rabiner & Huang, 1993). The NN model attempts to combine a large number of simple processing elements to simulate a biological system such as the human brain. The structure of neural networks makes it ideal for parallel computation. However, a disadvantage of NN is that the framework is static – it is difficult to handle the temporal structure of speech signals (Bimbot et al., 2003; L. R. Rabiner & Huang, 1993). Support Vector Machines (SVM) (Shawe-Taylor & Cristianini, 2000) is a relatively recent development in speech recognition and keyword spotting. However, no research by Australian institutions involving SVM has been found, so it will not be discussed here. Recently, more attention has been paid to extending NN, e.g. time delay recurrent NN (Weibel, Hanazawa, Hinton, Shikano, & Lang, 1989; Zhou, Liu, Song, & Yu, 1998) to address this issue. Before the nineties, HMM (L. Rabiner & Juang, 1986) was considered the dominant approach. The current trend is to combine HMM with NN to form so-called hybrid models (Trentin & Gori, 2003). Excellent surveys can be found in (Morgan & Bourlard, 1995; Trentin & Gori, 2001).

4.3.1 Self Organising Maps

Sehgal et al. from Monash University have developed `UbiqRec` (Sehgal, Gondal, & Dooley, 2004), a speech recognition system based on Self Organizing Maps (SOM) (Kohonen, 1993), which are a subtype of neural networks. It can be used for phoneme recognition with the number of output neurons equal to the number of phonemes to be recognized. However, Sehgal et al. showed that recognition performance can be significantly improved if multiple SOM are used, each SOM optimizing their weights for a specific phoneme class. This idea, known as Concurrent SOM, has also been used successfully in image processing applications such as multispectral satellite imagery (Neagoe & Ropot, 2004). The obvious disadvantage is that of increased computational complexity, but Sehgal et al. alleviated this problem with the use of Singular Value Decomposition (SVD). `UbiqRec` is novel in that it uses the Arabic language, which is rarely used in speech research. However, it is not in the public domain.

4.3.2 Hidden Dynamic Models

A relatively recent development in LVCSR is the use of Hidden Dynamic Models (HDM) to account for the weaknesses of HMM. It is well-known that HMM has difficulties with co-articulation and phonological variation, problems which are specific to LVCSR. The HMM is a data-driven approach which does not take into account the underlying human speech production process. It has a large number of parameters and is difficult to adapt to new speakers, without the use of unreasonably large amounts of training data. On the other hand, the HDM is a more structured model of speech production that respects the manner in which humans produce speech. This has been investigated by UWA. In (Togneri & Deng, 2004; Togneri & Li, 2001), Togneri et al. proposed that Vocal Tract Resonances (VTR) can be used as an alternative to MFCC since VTR space has lower dimensionality. Thus there are fewer parameters to estimate and less training data is required. The "learning" or the estimation of state and parameter information was based on the Expectation Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). The E-step required the calculation of conditional expectations (the sufficient statistics), which was then input into the M-step. The sufficient statistics were calculated using the Extended Kalman Filter (EKF). Togneri and Deng showed

that it is possible to use EKF for estimation of both state and parameters, instead of just state (Deng & Ma, 1999). They also showed that the HDM was capable of generating speech vectors that give good approximations to real data.

4.3.3 Cross Language Adaptation

QUT has a strong interest in developing speech recognition technology for the resource-poor Indonesian language (T. Martin, 2006). In (T. Martin & Sridharan, 2005; T. Martin, Svendsen, & Sridharan, 2003), Martin et al. proposed the use of cross-language adaptation: a recognizer trained in one resource-rich language can be used to adapt to a resource-poor language. Cross language adaptation has been studied before, however the earliest efforts focussed on a one-to-one mapping between phonemes of different languages. This leads to context mismatch: although the same phoneme is shared in different languages, they do not occur in the same context. For instance, the trigram SCH occurs frequently in German, but hardly in English (although this example is at the letter level, a similar phenomenon occurs at the phoneme level).

To solve this, Polyphone Decision Tree Splitting (PDTS) (Schultz & Waibel, 2000) was used. In PDTS a Context Querying Decision Tree is constructed for each phoneme, where each node represents the occurrence of any phoneme occurring in different contexts. Starting with a single node, the principle of maximum entropy gain was used to decide which nodes to split. The "formative branches" of the tree was built using the target language (this ensures the state distribution of the final model is compatible with the target language requirements). The training data from the source language was then used to extend the tree. Martin and Sridharan proposed a number of improvements for PDTS (T. Martin & Sridharan, 2005): for instance, phonemes were grouped according to whether they are vowel or consonant as well as state in a 3-state left-to-right topology, i.e. six decision trees are built per phoneme. Models were separately trained for noise, silence, pauses etc. Experiments were performed on Switchboard-I (SWB-1-ENG), 1996 HUB5 evaluation Spanish data (HUB5-SPAN) and Indonesian speech from OGI Multilanguage Speech corpus. The proposed method (named NEW-Tech) was compared with a knowledge driven mapping technique based on IPA combined with a context dependent model training paradigm (referred to as Know+STD). NEW-Tech outperformed Know+STD if adaptation data was used. However, it performed worse than a "baseline" system based on 2 hours of Indonesian speech (instead of 90 minutes or less for New-Tech and Know+STD). This result suggests that lack of data remains a serious difficulty in speech recognition. Martin et al. acknowledged that it is difficult to determine the merits of cross-language modelling due to the poor results. However, further investigation may lead to better performance.

4.4 Decoding

Decoding is the problem of choosing the word sequence with highest probability. This is a non-trivial problem since it is not feasible to search through every conceivable sequence of words. The problem is essentially that of searching a large solution space for an optimal solution with highest score and the standard AI techniques such as depth-first, breadth-first, beam-search and A* (Luger, 2002) have all been applied. The reader is referred to (S. Young,

1996) and references therein for more detail. No significant research by Australian institutions into the decoding phase of a speech recognition system has been found.

5. Keyword Spotting

5.1 Overview of KWS

Keyword spotting, also known as word spotting, is a "simplified" variant of speech recognition, where it is only necessary to detect certain words of interest rather than the entire speech utterance. This allows KWS systems to analyse an audio signal in less than real time, whereas speech recognition systems typically take much longer, especially in the context of LVCSR. KWS has many applications such as information retrieval from stored speech, detection of command words in voice operated software etc. Although KWS has become an active research area in recent years, it is not nearly as popular as speech recognition. Nevertheless there has been some interesting research in Australia.

There are many similarities between KWS and speech recognition. The performance decreases if the size of the vocabulary increases or the speech is fluent, as one would expect. For instance, (Lee, Soong, & Paliwal, 1996) cited an example where word spotting on a 5-word isolated keyword recognition task yielded much better performance than word spotting on a 20-keyword fluent speech keyword recognition task. As well, the modelling techniques of HMM, NN and SVM which have been successfully applied to speech recognition can also be applied to KWS. For instance, when using the HMM model, a common assumption is that explicit background "filler models" can represent all out-of-vocabulary or non-keyword speech (Higgins & Wohlford, 1985; Lee, Soong, & Paliwal, 1996; R. C. Rose & Paul, 1990). The features used in speech recognition such as MFCC's plus delta and acceleration coefficients are typically used for KWS as well. For these reasons, a separate chapter for parameterization and modelling will not be provided here, but individual contributions by Australian research institutions will be discussed immediately below.

5.1.1 Dynamic Match Phoneme Lattice Spotting

HMM-based keyword spotting suffers from very slow query speeds. In Phoneme Lattice Searching (PLS), Young et al. (S. J. Young, Brown, Foote, Jones, & Jones, 1997) attempted to improve query speeds by indexing speech files with a lattice. Each file was represented by a phoneme-lattice, which can be very efficiently traversed during query time. However, a serious drawback of PLS is that target phoneme sequences must either be detected as an exact match or rejected outright, thus yielding a high miss rate. Thambiratnam and Sridharan from QUT (Thambiratnam & Sridharan, 2005) addressed this drawback via Dynamic Match Phoneme Lattice Spotting (DMPLS). DMPLS effectively allowed for phoneme recognizer errors such as substitution deletion and insertion. More specifically, a sequence was accepted if the Minimum Edit Distance (Jurafsky & Martin, 2000) from the correct word was below a threshold. The word "dynamic" was derived from the well-known dynamic time warping principle (L. R. Rabiner & Huang, 1993), where the minimum number of insertions, deletions and substitutions to convert one sequence of phonemes to another is calculated. DMPLS was compared against a conventional HMM-based keyword spotting system (Rohlicek, 1995) using the Switchboard-1 conversational telephone speech corpus and the TIMIT microphone speech database. Although the miss rate of DMPLS was slightly inferior to that of HMM

(13.9% versus 8.0%), a tenfold improvement in both false alarm rate and computation speed was reported. Hence DMPLS is suitable for tasks that require high speed and accuracy.

5.1.2 Relationship between training data and performance

An important issue is how the amount of training data affects keyword spotting performance. In speech recognition, the lack of training data is a common complaint among many researchers, and there is a strong demand for the amount of training data to be increased, possibly by 1 or 2 orders of magnitude. But it is unclear how much performance gain this may yield. In fact Moore (Moore, 2003) claimed that an inordinately large amount of training data would be required to bring the performance of an automatic speech recognition system equal to that of a human listener.

It is expected that a lack of training data would also decrease performance in KWS. However, KWS is a much more constrained task, attempting to discriminate between a small set of classes. Hence it is reasonable to hope that KWS may be less affected by reduced amounts of training data than speech recognition. If so, KWS techniques may provide a viable short-term solution for the development and deployment of non-English data mining applications. In (Thambiratnam, Martin, & Sridharan, 2004), Thambiratnam et al. investigated the effect of limited training data on the performance of KWS systems. Experiments and discussion were presented to assess the benefits of a large training corpora for KWS, and to determine whether the benefits form the increased training data provided sufficient gains to motivate the collection of this data. The languages examined in this study were English, Spanish and Indonesian. Encouragingly, the research indicated that KWS is significantly less sensitive to the training databases size, when compared to speech transcription. For example, it was found that reducing 160 hours of training data for English to 4 hours resulted in only a 6.1% loss in EER[7], which is significantly less than the 18% reported by Moore for speech recognition. Similar results were observed for Spanish and Indonesian.

[7] In keyword spotting the EER is typically used, rather than the word error rate for speech recognition.

6. Accent Identification

6.1 Overview of Accent Identification

AID is closely related to language identification, except that all speakers speak the same target language. It is expected that speakers with foreign accents will import some aspects of their first language when speaking the target language, and it is well-known that speech recognition systems significantly degrade when the speaker accent differs from that in the training set. Thus knowledge gained from accent ID can often improve speech recognition performance. It is also an important tool for forensic applications. The components are similar to speaker or language identification (see Figure 6):

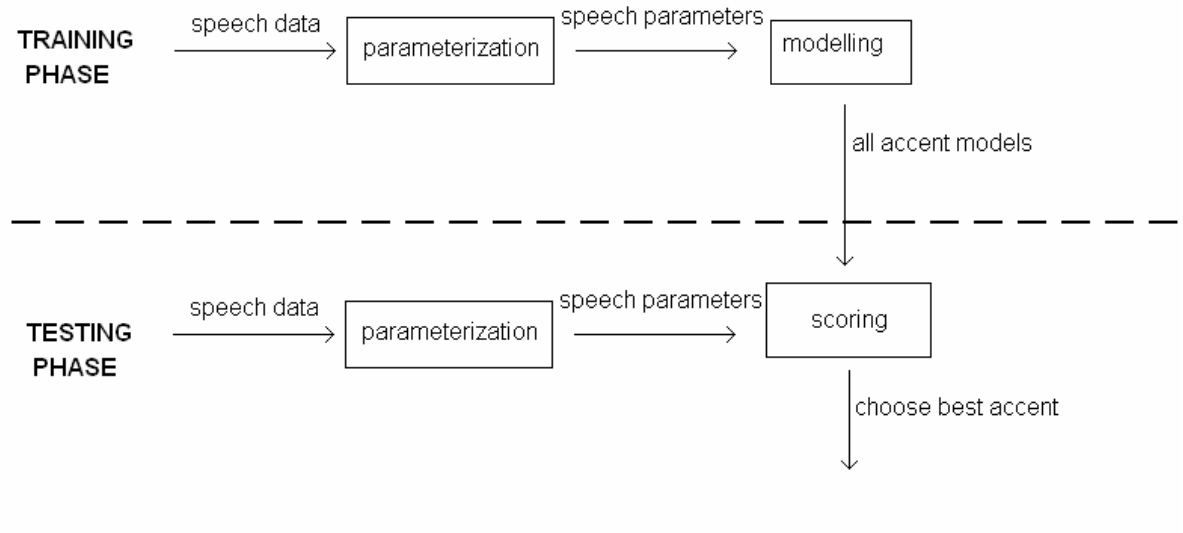

Figure 6: Modular representation of accent identification system

6.2 Parameterization

There are a number of possibilities that have been considered regarding feature space for accent ID: prosody, sub-word, spectral and word-based approaches have all been proposed (Tanabian & Goubran, 2005). Prosody is a discriminative feature in foreign accent identification. Hansen et al. (Hansen, Yapanel, Huang, & Ikeno, 2004) used normalized fundamental frequency (F0) range and syllable rate to distinguish different accents. Spectral features such as MFCC, delta MFCC, log energy and delta energy have been used (Kumpf & King, 1996). They do not offer any specific advantages, except that these spectral features are ubiquitious in all areas and applications of speech processing. Sub-word modelling is another approach. The idea is to detect phonemes from a different language used to approximate the "correct" phonemes in the target language, typically using PPRLM (Zissman, 1995; Zissman & Singer, 1994) which has achieved state-of-the-art performance for both language and accent

identification. The main advantage of word-based approaches (Rongqing & Hansen, 2005) is that an entire word carries much more information than a phoneme. Hence it is possible to exploit dialectal variations when different speakers utter the same word. Sub-words are easier to work with than whole words, largely because of the small number of the former compared to the latter. However, most speech (e.g. day-day conversation) only uses a small fraction of all possible words in a language (Tanabian & Goubran, 2005). By only considering the most common words, this approach avoids the "curse of dimensionality" that one would expect from an entire vocabulary. The choice of parameterization remains an unsolved problem, typically "solved" using an arbitrary choice depending on the application.

There are a few Australian papers on AID. The following paper, a joint effort from MIT and University of Sydney, (Berkling, Zissman, Vonwiller, & Cleirigh, 1998) used the subword approach. Past research focussed on phoneme inventories, phoneme sequences and intonation patterns. In this paper a new feature was proposed: *location of phoneme within a syllable*. It is well-known that syllables can be subdivided into onset and rhyme, but this does not indicate where a syllable occurs within a word. Berkling et al. defined three constituents: proclitic, core and enclitic. The core contained the obligatory vowel. The proclitic and enclitic constituents covered components that only occurred morpheme-initially and morpheme-finally respectively, and indicated a boundary of grammatical units in English. For each accent a confusion matrix was computed relating the probability of a target phoneme given an achieved phoneme. For an achieved phoneme sequence, the accent was classified according to the confusion matrix which best "explained" the difference between the target and achieved phoneme sequences. Knowledge of English syllable structure could be incorporated by treating the confusion matrix as a function of position (proclitic, core, enclitic) rather than "constant" (with respect to position). The modelling component used the standard HTK to recognize 40 different phoneme models. Berkling et al. hypothesised that identifying Lebanese accents is harder than identifying Vietnamese, since the pronunciation of the former is closer to native English than the pronunciation of the latter. They tested their algorithm on two- and three-way classification of English (EN), Vietnamese (VI) and Lebanese (LE) accents. For two-way accent identification an improvement from 86% to 93% was achieved for EN-VI and an improvement from 78% to 84% was achieved for EN-LE. These results support their hypothesis. For three-way accent classification the accuracy improved from 69% to 77%.

In (Kumpf & King, 1997), from Speech Technology Research Group (STRG), UTS[8], Kumpf and King did not propose a specific set of features but instead attempted to associate a different feature set for different phonemes. Thus the set of features was not "constant" but was rather a function of phoneme. A single feature vector was extracted for all phoneme classes combining acoustic (MFCC, log energy), prosodic (segment duration, F0, delta F0) and contextual information (description of phonemic left and right context). For each phoneme class Linear Discriminant Analysis (LDA) was used to select a different subset of the above features. This was used to keep features that assisted in accent discrimination but eliminated redundant features that did not contribute much to accent discrimination capability. Kumpf and King claimed that their accent classification scheme achieved performance close to the human benchmark.

[8] The authors were working at Sydney at the time of writing but King is currently residing at University of South Australia

6.3 Modelling

The simplest model that has been used is GMM[9] (Too, Chao, Chang, & Jingehan, 2001). The main advantage is that it requires no segmentation or phonemic labelling of training speech. However, it cannot model temporal information of the speech signal. Neural networks have been used only rarely (Tanabian & Goubran, 2005) and will not be considered in detail here. The HMM is the most common model since it accounts for both temporal and spectral variations in the speech signal. However, it does require phonemic labelling or segmentation of training speech. Phonemic labelling can be done either manually or automatically. Manual labelling or segmentation is time-consuming and the results are often inconsistent, even among experts. However it is more accurate than automatic labelling or segmentation. This can be an important consideration when AID is part of a larger system (e.g. speech recognition) since errors in one module can propagate to the next. A related issue is the fact that manual labelling is often only available for a small percentage of languages, so experiments are typically restricted to a corresponding subset of a particular database. For instance, in (Kumpf & King, 1996) (see below), only three accents of English were considered when there are three varieties of Australian English (General, Broad and Cultivated) and ten foreign-accented varieties of English in total.

Unfortunately, AID is a relatively new field and research results are limited, even outside Australia. Some research has been done within a single model, such as exploring the effect of the number of components in a GMM (Too, Chao, Chang, & Jingehan, 2001), but no important comparisons between different models have been found. In LID, it is well established that GMM is inferior to other approaches such as PPR (Zissman, 1996), but there is no corresponding comparison in AID.

Kumpf and King (Kumpf & King, 1996), from STRG, UTS, used a Parallel Phoneme Recognition (PPR)-based system (Hazen & Zue, 1993; Zissman, 1996; Zissman & Singer, 1994) for automatic accent classification for foreign accents. They used accent specific HMMs and phoneme bigram language models to derive accent discrimination likelihood scores. Speech was automatically segmented using a HMM segmenter trained on Australian English (AuE) phoneme classes. This was used to train accent-specific HMM-phoneme and phoneme-bigram models. The PPR approach is as follows: during training, a HMM phoneme model and language model (phoneme bigram) are used on three accents, namely: Australian English, Lebanese Arabic and South Vietnamese. During testing, a Viterbi decoder determines the most likely state sequence representing the speech utterance, given the HMM and language model corresponding to an accent. Log likelihood scores are assigned to the state sequences. A bias is subtracted to account for the small training size of the database. Finally, the accent corresponding to the highest final score is determined as the "correct" accent. This process is demonstrated in Figure 7.

[9] Unlike the usage of GMM in speaker identification, this paper is not from an Australian research institution.

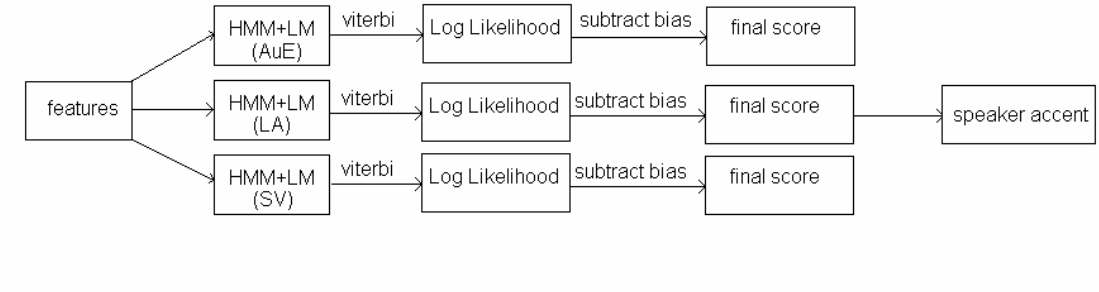

Figure 7: Modelling component of accent identification system

In the experiments of Kumpf and King the ANDOSL database was used (Millar, Vonwiller, Harrington, & Dermody, 1994). They tested their algorithm on two- and three-way classification of Australian, Lebanese-Australian and South Vietnamese accents. For the two- and three-accent classification tasks, the best average obtained was 85.3% and 76.6% respectively. The use of the language model (as opposed to using only the HMM) only contributed marginally to overall performance (about 2-3%). A study was also performed on the effect of the use of automatic segmentation and it was found that manual segmentation only yielded an improvement of 1-2% to the overall score. Kumpf and King postulated that the difficulties of manual segmentation and consequent lack of training data had caused the use of phoneme-bigram language models and manual segmentation scoring only marginal improvement in their experiments.

The ANDOSL database is very significant in the context of language/accent identification since it arose out of a project funded by the Australian Research Council, involving a number of research groups from various universities. This will be discussed in more detail in section 9.

7. Phoneme Recognition

7.1 Overview of Phoneme Recognition

Phoneme recognition is not really a "functional component" that is useful by itself. It is nevertheless an important problem as it is frequently used in other components such as speech recognition. In fact, phoneme and speech recognition have many similarities. For instance, front end processing techniques are similar and it is common to model phonemes using a left-to-right HMM with three states. One important and obvious difference is that evaluation is performed at the level of phonemes instead of words (e.g. in speech recognition the Word Error Rate is often used). A corollary is that phoneme recognition does not require the use of word models as described in section 4.3, so it is somewhat simpler. Unfortunately, no significant reviews on phoneme recognition have been found, and there are few papers describing phoneme recognition by itself rather than speech recognition. Therefore this report will only briefly discuss three papers from QUT, Newcastle and NAL, and Edith Cowan University.

7.2 Phoneme Recognition with LID

In (Wong & Sridharan, 2003), from QUT, Wong and Sridharan examined the problem of multilingual phoneme recognition, where it is necessary to both identify the language and generate a phoneme sequence. They considered three different approaches to multilingual phoneme recognition, which were labelled Approach 1, Approach 2, Approach 3 (A1,A2,A3). Essentially language identification can be done either explicitly or implicitly (the first two methods A1 and A2 are explicit). In A1 each language was modelled by a GMM, as done in speaker verification. The language that best matches the speech (in the sense of highest likelihood score) was identified and the monolingual system corresponding to that language was used to perform phoneme recognition. The GMM is an unusual choice for the modelling component, since, unlike most other models, it does not employ phoneme recognition at any level. Alternatively, in A2 all monolingual systems (corresponding to all languages) were used to perform phoneme recognition and the highest likelihood score determined the phoneme transcription, as well as the language. This avoided the use of GMM in A1, but was computationally more expensive since all monolingual systems must be employed instead of only one in A1. The implicit LID method (A3) mapped phonemes from multiple languages into a smaller multilingual set of phonemes. This allowed the recognition system to handle utterances in multiple languages. However language information was lost, which mitigated against certain language-specific speech recognition techniques. Given a choice of approach A1, A2 or A3, phonemes were modelled using a 3-state HMM with 8 Gaussian mixture components per state. Wong and Sridharan also defined a "baseline" algorithm which assumes that perfect language identity information is always available. Thus the baseline was expected to be better than A1, A2 or A3.

In the experiments A1, A2, A3 and the baseline algorithms were tested on both isolated and continuous speech in three languages, namely English, Mandarin and Spanish. Wong and Sridharan concluded that the superiority of one method over another largely depended on the

LID stage. Given high LID accuracy, explicit-LID was better than implicit-LID. More specifically, A1 was the best of the three methods but was still 5% worse than the baseline performance. If the LID accuracy was poor, as in the isolated phoneme recognition experiment, then A3 was the best of the three methods, but was still much worse than baseline.

7.3 Phoneme Recognition Using Wavelet Transforms

Tan et al. from the University of Newcastle and National Acoustic Laboratories, NSW, (Tan, Minyue, Spray, & Dermody, 1996) used the wavelet transform as a front-end pre-processor for HMM-based phoneme recognition. Two versions of the wavelet transform were tested, namely, the Discrete Wavelet Transform (DWT) and Sampled Continuous Wavelet Transform (SCWT). The use of Mel-scale cepstral coefficients of order 12 was also tested and served as the baseline. The main advantage of SCWT is its ability to preserve both harmonic and formant information from the speech signal. However, results on the prototype version (1988) of the TIMIT database suggested that SCWT was only marginally better than the baseline in recognition rate, but DWT was significantly worse.

7.4 Time-Frequency Shift-Tolerant Pre-processing

As remarked earlier, a significant problem of Neural Networks (NN) is that they are often unable to recognise a time shift in the input speech signal. Hence it is useful to find an input pattern that is independent of any time-shift relative to the training pattern. Similarly, one often wants to be able to recognise a frequency shift in the input speech signal (for instance, "helium speech" can be simulated by a large amount of frequency shift and a high-pass filter). Several researchers have proposed different NN architectures to obtain both time-shift and frequency-shift invariance in the input pattern for various applications including speech recognition (Sawai, 1991) and phoneme recognition (Basu & Svendsen, 1993). In (Ang & Hon Nin, 1995), from Edith Cowan University, Ang and Hon Nin used a spectrogram as the time-frequency distribution and a counter-propagation network (Hecht-Nielsen, 1988) to recognise a fixed-position pattern. Given a spectrogram, a two-dimensional FFT was used to obtain a time-frequency shift-tolerant pattern. This was then input to a counter-propagation network. Ang and Hon Nin showed that their algorithm can distinguish between the five different vowels of English: 'a', 'e', 'i', 'o', 'u'.

8. Speech Segmentation

8.1 Overview of Speech Segmentation

Speech segmentation is the problem of locating boundaries between sounds corresponding to the phonemes that make up a speech signal. It is possible to perform segmentation manually, especially in applications where the highest precision is critical. However, more research is being performed on automatic segmentation, which is useful for applications where precision is not critical. For example, automatic segmentation is sufficient when training HMM for speech recognition since segmentation errors are "averaged out" (Cox, Brady, & Jackson, 1998).

Usually the phoneme sequence is given, in which case the segmentation problem is also known as forced alignment. Although speech segmentation is not directly useful for an end-user, it has many important applications. For example, for a corpus to be useful for speech recognition research, the speech itself should be complemented with phoneme labels and segmentation. Since speech segmentation is not directly useful for an end-user, it is not clear how best to measure the quality of automatic segmentation. The most common measure is the percentage of boundaries that are correctly located, to within a specified tolerance, e.g. "96% of boundaries with errors below 20ms".

Automatic speech segmentation consists of parameterization to obtain feature vectors and the usage of different "techniques" to obtain the final segmentation (see Figure 8). The word "modelling" is deliberately avoided since, unlike some of the other fields (e.g. speaker recognition), the decision may not depend on a specific model. In an extreme case, (Alani & Deriche, 1999) , which will be discussed below, the segmenting technique is nothing more than thresholding a distance measured between four contiguous frames.

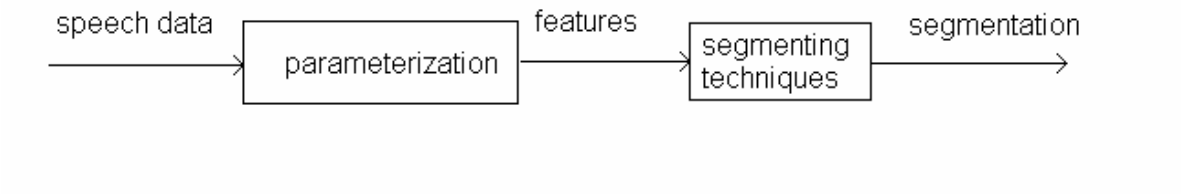

Figure 8: Modular representation of speech segmentation system

8.2 Parameterization

There are a number of choices for the parameterization stage. For instance, the f0 contour, short-time energy contour and energy in different frequency bands have all been used. The reader is referred to Toledano et al. and references therein for a more detailed discussion (Toledano, Gomez, & Grande, 2003).

The wavelet transform is a powerful tool in signal analysis. It is probably best known for its application to image compression in JPEG, but it has also proved useful for speech compression (Agbinya, 1996). The wavelet transform is used to analyse a signal in both time and frequency domains. By using both short high-frequency and long low-frequency windows, the wavelet transform can be used to detect fast transients (stops) and slow transients (vowels).

In (Alani & Deriche, 1999), from QUT, Alani and Deriche used the wavelet transform for speech segmentation. At any point in time the presence or absence of a boundary was determined by considering the values of the feature vectors in four contiguous frames. If the first two frames were sufficiently different from the latter, a boundary was detected. Experiments have been performed on the TIMIT database. An unusual aspect of the experiments was that the wavelet transform was compared against spectrum coefficients instead of the more common cepstral coefficients. The wavelet transform was more compact, using six parameters instead of sixteen in the Mel-scale spectrum coefficients, and the performance was slightly superior in terms of both accuracy and false alarm rate. Unfortunately the wavelet transform has rarely been used in speech segmentation and has not shown any great success.

8.3 Modelling

Three common modelling techniques are HMM, neural networks and dynamic time warping. HMM is the most common model since it has already been extensively studied in other areas of speech recognition. Typically, a HMM-based phoneme recognizer would be altered by incorporating the known phoneme sequence which can be used for forced alignment. “Hybrid” approaches have also been proposed, where HMM is combined with other models and techniques. Again it is outside the scope of this report to discuss these in detail, and the reader is referred to (Toledano, Gomez, & Grande, 2003) and references therein.

9. Other Topics

9.1 Overview

This section covers a number of miscellaneous problems in speech processing, namely speaker diarisation (meeting segmentation) and database annotation. These "lesser-known" problems are not necessarily less important than those described in previous chapters and it is conceivable that more research could be devoted to them.

9.2 Speaker Diarisation

Speaker diarisation is the problem of segmenting an input audio channel into speaker 'turns' and associating a speaker label with each turn. A speaker diarisation system from Macquarie University (Cassidy, 2004a, 2004b) was developed primarily for participation in the NIST RT04 Spring evaluation. The system comprised of four components: (i) speech/silence classification, (ii) speaker segmentation, (iii) clustering and (iv) identification (see Figure 9).

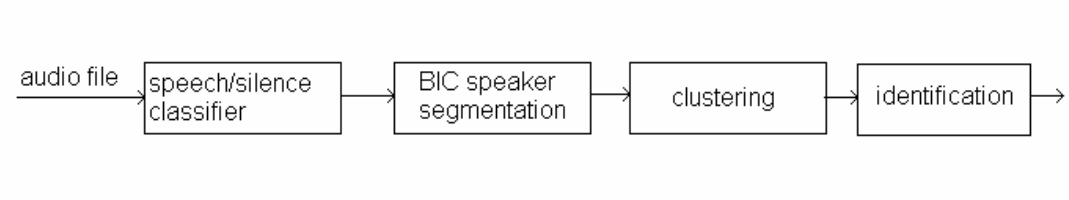

Figure 9: Modular representation of a simple diarisation system

The speech/silence classifier was used to remove the portions where no speaker was present. To detect speaker changes within a portion of speech, the popular Bayesian Inference Criterion (Chen & Gopalakrishnam, 1998) was used. This amounts to declaring that a speaker change occurs if there is a qualitative change in acoustic signal. Clustering was then performed to determine the number of speakers in the meeting, this information being unavailable in the RT04 Spring evaluation specification. Finally speaker models, derived from simple Gaussians (essentially these are GMM with the number of mixtures equal to one), were used to identify speakers.

Macquarie University's system was relatively simple and was not competitive in the NIST evaluation. Other more powerful systems have additional components. For instance, gender (male/female) or bandwidth (low/high) classification have been used to assist the clustering stage. A re-segmentation stage was often performed to refine original segment boundaries and fix "small errors". An excellent review can be found in (Tranter & Reynolds, 2006).

9.3 Database Annotation

Although large speech databases have been constructed world-wide (see for instance (Lai et al., 1990) references 1-8), this is somewhat neglected in Australia. One of the research efforts in

Australia at creating a database is due to UWA. Lai et al. (Lai et al., 1990) have developed a simple database, unfortunately unnamed, containing speech from Australian speakers of various ethnic groups. The speakers consisted of equal numbers of males and females and most of them had tertiary education. Recording was performed in a quiet room with minimal noise. The speech material was divided into four categories: isolated digits/numbers/words, vowels/diphthongs, connected digits and six phonemically balanced sentences. At the time of writing, voices from 110 subjects have been collected, and Lai et al. are intending to collect more data.

An important database is ANDOSL, which was briefly mentioned in section 6.3. As it is intended primarily for research only within Australia, the database comprises of spoken language from a variety of major speaker groups in Australia. Although four institutions (Sydney, NAL, Macquarie, and ANU) have cooperated in the ANDOSL project, only Macquarie University performed database annotation (which was identified as one of the "components" of the ANDOSL project). Nowadays large databases of speech data are readily available and researchers have focussed on development of annotation standards. Clearly, it is desirable to develop software tools and labelling standards for general purposes, instead of just one or two ad-hoc specific projects, such as in (Lai et al., 1990). Cassidy and Harrington (Cassidy & Harrington, 1996; Cassidy & Harrington, 2001) from Macquarie University have developed EMU[10], a speech data management system designed for labelling and querying databases such as ANDOSL or TIMIT. EMU is an extension of a previous system called MU+ (Harrington, Cassidy, J., & A., 1993). It is a flexible system and offers many advantages over previous ad-hoc systems. For instance, it has databases for languages other than English, and EMU can read a number of popular label and data formats.

The fundamental principle of EMU is combining both sequential and hierarchical information. The term "hierarchical" means that information can be represented at different levels, such as phonemes, words or complete sentences. For instance, if a child whispers a particular sentence it is natural to represent it at the utterance level, rather than insert a "special token" at the phonemic level to represent the beginning or ending of a whispered sentence (Cassidy & Harrington, 1996). Thus the information that can be represented is much richer and more powerful than that represented by the sequential information alone. EMU was implemented as a C++ library. Graphical user interfaces to various database functions could be implemented with the help of the Tcl/Tk scripting language. Using Tcl, EMU could be converted from a hierarchical database to a relational database. Cassidy showed that this can improve query speeds, especially for large databases (Cassidy, 1999). Originally based within SHLRC at Macquarie University, the EMU project has now been developed into an international collaboration (http://emu.sourceforge.net/).

[10] EMU is merely a name, not an acronym.

10. Summary and Conclusions

The previous chapters give an overview of the kind of work done by various institutions. It should be clear that there is a large variety of work done within Australia, in terms of task complexity, relevance to defence applications, amount of effort invested and quality of research. This report has included a significant number of research papers in Australia to give a rough snapshot of where each individual research institution stands. The purpose of this chapter is to compare the relevance of the work done by each institution in greater detail.

Although research institutions often have interests in a particular functional component, such as speaker identification, it is equally likely to research on a particular operational component, and apply it to different functional components. Therefore two tables have been displayed. The first table compares institutions versus functional components:

Table 1: Functional components used in Australian research institutions

RESEARCH INSTITUTION	FUNCTIONAL COMPONENTS
ANU	SID
EDITH COWAN UNIVERSITY	Phoneme recognition
GRIFFITH	SID + speech recognition
MACQUARIE	Speaker diarisation, database annotation
MONASH	Speech recognition
NEWCASTLE+NAL	Phoneme recognition
NICTA	SID + LID + speech recognition
RMIT	SID
QUT	SID + LID + speech recognition + segmentation
SYDNEY	Speech recognition + accent detection
UNSW	SID + LID + speech recognition
UWA	SID + speech recognition + database annotation

For ease of presentation of this table, SID is used to refer to both speaker identification and verification. Similarly, speech recognition covers both keyword spotting and continuous speech recognition. As can be seen, SID and speech recognition are the most popular areas. The contribution by ANU into SID is particularly important since Rose et al. have done a significant amount of research into forensic SID (P. Rose, 2006; P. Rose, Kinoshita, & Alderman, 2006; P. Rose, Osanai, & Kinoshita, 2003). QUT has done a significant amount of work on speech recognition, including some studies on the interesting topic of Audio-Visual Speech Recognition (AVSR) (Dean, Lucey, & Sridharan, 2005; Dean, Lucey, Sridharan, & Wark, 2005). Although speech synthesis is also popular in Australian research institutions, it is omitted from the table as it is irrelevant for the purpose of this report.

Not all the "big universities" are present in the table. Notable absentees include the University of Adelaide, University of South Australia and University of Melbourne. No significant research in these institutions has been found. Although Melbourne has done some work on

speech processing, their work mostly involved language development for disabled people with severe or profound deafness, and for that reason they have not been considered. In contrast, Griffith University, which is not a Group-Of-Eight member, has done some relevant research in SID and speech recognition.

The second table summarises which institutions have used which operational components and which have done research into which ideas. It does not necessarily indicate which idea was *invented* by which institution.

Table 2: Operational components used in Australian research institutions

RESEARCH INSTITUTION	**OPERATIONAL COMPONENTS**
ANU	Diphthongs and F-patterns
EDITH COWAN UNIVERSITY	Neural networks
GRIFFITH	Phase information, subband spectral coding histogram
MACQUARIE	GMM (diarisation), hierarchical database annotation
MONASH	Self organizing maps
NEWCASTLE+NAL	Wavelet transform
NICTA	Phase information, prosody and spectrum information, feature warping
RMIT	Discriminative feature extraction
QUT	GMM, feature warping, LPCC vs MFCC, wavelet transform, hybrid systems
SYDNEY	Location of phoneme within syllable
UNSW	Phase information, prosody and spectrum information
UTS	Syllable structure (in AID), different feature sets per phoneme
UWA	Trajectory models, hidden dynamic models

Although some papers described an entire system, e.g. (Cassidy & Harrington, 1996) most described only a single idea or technique such as utilizing phase information (e.g. Modified Group Delay Coefficients) in the context of a specific functional component (e.g. speech recognition). It was also common for one research institution to exploit research done somewhere else, which is especially true for a single operational component being used in multiple contexts. For instance, since feature warping from QUT has proved to be so successful, it has been regularly used by several research institutions in many areas of speech processing. NICTA used feature warping (also known as cumulative distribution mapping) in the context of noise-robust speech recognition although that was only for digit recognition. Also, QUT and NSW have used the wavelet transform for speech processing merely because it is a well-known concept in signal processing in general, not because of any special advantages of the wavelet transform in the context of speech processing.

Comparison between different items of research is difficult for many reasons. There are a large number of different corpora for all functional components. Even when two parties work on the same functional component and the same corpus, it is common for at least one side to restrict experiments to a subset of a particular corpus. Moreover, the success of an experiment

is by no means the only indication of how useful a particular direction of research is. For instance, other important factors include the placing of software in the public domain or participation in standard evaluations such as the well-known NIST evaluations.

The NIST evaluations are designed to advance the state of the art in various tasks such as speech recognition, speaker recognition and language identification etc. Although the evaluations aim to simulate realistic conditions (e.g. telephone conversations where channel mismatch between training/test data is common), they are not a 'perfect' indicator of real-world performance. For instance, a system for speaker recognition may perform excellently in a NIST evaluation but do poorly in a real-world application. Nevertheless, these evaluations are highly respected in the speech processing community as they enable researchers to directly compare different algorithms using both the same test data and same evaluation specification plan. Moreover, the tasks are complex and the testing is quite thorough. Typically, participation in a NIST evaluation requires the submission of "complete results" for one or more test conditions. Good performance in a NIST evaluation generally indicates that a system is highly competitive with other systems worldwide.

Unfortunately, most of the research papers listed in this report have avoided the NIST evaluations, either by using an "inferior" corpus (corpora) or by ignoring the specification plan for the given corpus (corpora). There are a number of reasons for this:

1) For a system to be competitive in a NIST evaluation, it requires the institution to invest many years of research on a specific task. This is a significant undertaking, even for the best research institutions in Australia.
2) Many applications only require simple tasks e.g. a typical telephone banking application only requires the recognition of ten digits plus a few simple words.
3) NIST evaluations are "closed shop". Only those who register or submit a system for evaluation are allowed to have detailed knowledge about all systems submitted, i.e. how they work, how they scored and which systems performed best overall. It is illegal for NIST participants to comment publicly on the relative performance of other participants. Hence it is difficult to obtain meaningful comparisons between state-of-the-art systems.

QUT is the only regular Australian NIST participant. Their papers listed in this report almost always conform to NIST specifications. Outside QUT some Australian research efforts use data derived from NIST (Allen, Ambikairajah, & Epps, 2006; Price, Willmore, Roberts, & Zyga, 2000), but these are exceptions rather than the rule. Macquarie University is the only other institution to have participated in a NIST evaluation, but their results were not competitive.

Another significant project is the creation of the ANDOSL database. Four universities contributed to this database: Sydney, NAL, Macquarie and ANU. The database has been used by Macquarie University for the tasks of database annotation and speaker diarisation (Cassidy, 2004b; Cassidy & Harrington, 1996). This has been described in section 9.3.

Some effort has been made in getting software into the public domain. The research on front-end processing performed by UWA includes a software system called `fview` which has been placed in the public domain (Tey, Jong, & Togneri, 1996). Although it has been used for

speech and speaker recognition, its primary objective is to promote research specifically for front-end processing regardless of application. However, it is not popular outside UWA. The EMU system from Macquarie (Cassidy & Harrington, 1996) is not only publicly available, but has also achieved international recognition.

A lot of Australian research into speech recognition involves only small vocabulary. As systems become more powerful, it is likely that more research will be focussed on more difficult tasks. It is encouraging that some Australian institutions are beginning to research difficult tasks. For instance, Togneri et al. (Togneri & Deng, 2004; Togneri & Li, 2001) from UWA and Thambiratnam from QUT (Thambiratnam & Sridharan, 2005) have done research on LVCSR/keyword spotting.

11. Acknowledgements

The author of this report would like to thank Terrence Martin, Jonathan Willmore, Richard Price and David Parker for technical discussions.

12. REFERENCES

Agbinya, J. I. (1996). *Discrete wavelet transform techniques in speech processing*. Paper presented at the Proceedings. 1996 IEEE TENCON. Digital Signal Processing Applications.

Alani, A., & Deriche, M. (1999). *A novel approach to speech segmentation using the wavelet transform.* Paper presented at the Proceedings of the Fifth International Symposium on Signal Processing and its Applications.

Alderman, T. (2005). *Forensic Speaker Identification : A Likelihood Ratio-based Approach Using Vowel Formants* Lincom GmbH.

Allen, F., Ambikairajah, E., & Epps, J. (2006, May). *Warped Magnitude and Phase-Based Features for Language Identification.* Paper presented at the International Conference on Acoustics, Speech and Signal Processing.

Ang, L.-M., & Hon Nin, C. (1995). *Counterpropagation network and time-frequency shift-tolerant preprocessing for phoneme recognition.* Paper presented at the 1995 IEEE International Conference on Neural Networks.

Atal, B. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *J. Acoust Soc Am., 55*(6), 1304-1322.

Auckenthaler, R., Carey, M., & Lloyd-Thomas, H. (2000). Score normalization for text-independent speaker verification systems. *Digital Signal Processing, 10*(1), 42-54.

Avendano, C., Van Vuuren, S., & Hermansky, H. (1996). *Data based filter design for RASTA-like channel normalization in ASR.* Paper presented at the International Conference on Spoken Language Processing.

Bacchiani, M. (2001). *Automatic transcription of voicemail at AT&T.* Paper presented at the International Conference on Acoustics, Speech, and Signal Processing.

Bahl, L. R., Brown, P. F., de Souza, P. V., & Mercer, R. L. (1993). Estimating HMM parameters so as to maximize speech recognition accuracy. *IEEE Trans. Speech and Audio Processing, 1*(1), 77-83.

Baker, B., & Sridharan, S. (2006). *Speaker verification using hidden Markov models in a multilingual text-constrainted framework.* Paper presented at the Odyssey 2006.

Baker, B., Vogt, R., & Sridharan, S. (2005). *Gaussian mixture modelling of broad phonetic and syllabic events for text-independent speaker verification.* Paper presented at the European Conference on Speech Communication and Technology, Lisbon.

Basu, A., & Svendsen, T. (1993). *A time-frequency segmental neural network for phoneme recognition.* Paper presented at the Proceedings of the International Conference on Acoustics, Speech, and Signal Processing.

Berkling, K., Zissman, M. A., Vonwiller, J., & Cleirigh, C. (1998, December 1998). *Improving accent identification through knowledge of English syllable structure.* Paper presented at the Proceedings of the 1998 International Conference on Spoken Language Processing, Sydney, Australia.

Bernard, J. (1967). *Some measurements of some sounds of Australian English.* Sydney University.

Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., et al. (2003). A tutorial on text-independent speaker verification. *Journal on Applied Signal Processing, 4*, 430-351.

Bo, Y., Ambikairajah, E., & Fang, C. (2006, 20-24 Aug). *Combining cepstral and prosodic features in language identification.* Paper presented at the 18th International Conference on Pattern Recognition.

Cassidy, S. (1999). *Compiling multi-tiered speech databases into the relational model: experiments with the Emu system.* Paper presented at the Proceedings of Eurospeech.

Cassidy, S. (2004a, December). *Evaluation of the Macquarie meeting room speaker diarisation system.* Paper presented at the Proceedings of the 10th Australian International Conference on Speech Science and Technology, Sydney.

Cassidy, S. (2004b, May). *The Macquarie speaker diarisation system for RT04S.* Paper presented at the ICASSP 2004 Meeting Recognition Workshop, Montreal, Canada.

Cassidy, S., & Harrington, J. (1996). *Emu: an enhanced hierarchical speech data management system*.

Cassidy, S., & Harrington, J. M. (2001). Multi-level annotation in the Emu speech database management system. *Speech Communication, 33*, 61-77.

Chen, S. S., & Gopalakrishnam, P. S. (1998). *Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion.* Paper presented at the Proc 1998 DARPA Broadcast News Transcription and Understanding Workshop, Landsowne, VA.

Choi, E. H. C. (2006). *A noise robust front-end for speech recognition using Hough transform and cumulative distribution mapping.* Paper presented at the International Conference on Pattern Recognition.

Cox, S., Brady, R., & Jackson, P. (1998). *Techniques for accurate automatic annotation of speech waveforms.* Paper presented at the Proceedings of the International Conference on Spoken Language Processing, Sydney, Australia.

Dean, D., Lucey, P., & Sridharan, S. (2005). *Problems associated with current area-based visual speech feature extraction techniques.* Paper presented at the Proceedings of the Auditory-Visual Speech Processing International Conference, Canada.

Dean, D., Lucey, P., Sridharan, S., & Wark, T. (2005). *Comparing audio and visual information for speech processing*.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society 39*(1), 1-38.

Deng, L., & Ma, J. (1999). *A statistical coarticulatory model for the hidden vocal-tract-resonance dynamics.* Paper presented at the Conf. EUROSPEECH'99.

Donglai, Z., & Paliwal, K. K. (2004). *Product of power spectrum and group delay function for speech recognition*.

Furui, S. (2000). *Digital speech processing, synthesis and recognition* (2nd ed.). New York: Marcel Dekker Inc.

Gajic, B., & Paliwal, K. K. (2006). Robust speech recognition in noisy environments based on subband spectral centroid histograms. *IEEE Trans. Speech and Audio Processing, 14*(2), 600-608.

Hansen, J. H. L., Yapanel, U., Huang, R., & Ikeno, A. (2004). *Dialect analysis and modeling for automatic classification.* Paper presented at the International Conference on Spoken Language Processing, Jeju Island, South Korea.

Harrington, J. M., Cassidy, S., J., F., & A., M. (1993). The Mu+ system for corpus based speech research. *Computer Speech and Language, 7*, 305-331.

Hazen, T. J., & Zue, V. W. (1993). *Automatic language identification using a segment based approach.* Paper presented at the Eurospeech 93, Berlin, Germany.

Hecht-Nielsen, R. (1988). Applications of counterpropagation networks. *Neural Networks, 1*, 131-139.

Hegde, R. M., & Murthy, H. A. (2005). *Automatic language identification and discrimination using the modified group delay feature.* Paper presented at the 2005 International Conference on Intelligent Sensing and Information Processing.

Hermansky, H., & Morgan, N. (1994). RASTA processing of speech. *Speech and Audio Processing, IEEE Transactions on, 2*(4), 578-589.

Higgins, A., & Wohlford, R. (1985). *Keyword recognition using template concatenation.* Paper presented at the Proceedings of the International Conference on Acoustics, Speech, and Signal Processing.

Jurafsky, M. J. H., & Martin, J. H. (2000). Speech and language processing: an introduction to natural language processing, computational linguistics and speech recognition, chapter Minimum Edit Distance. In: Prentice Hall.

Kohonen, T. (1993). Physiological interpretation of the self-organizing map algorithm. *Neural Networks 6*, 895-905.

Kumpf, K., & King, R. W. (1996). *Automatic accent classification of foreign accented Australian English speech.* Paper presented at the Proc. ICSLP'96.

Kumpf, K., & King, R. W. (1997). *Foreign speaker accent classification using phoneme-dependent accent discrimination models and comparisons with human perception benchmarks.* Paper presented at the Proc. Eurospeech97.

Lai, E. M., Carrijo, G. A., Bennett, R., Togneri, R., Alder, M., & Attikiouzel, Y. (1990). *An English language speech database at the University of Western Australia.*

Lee, C.-H., Soong, F., K. , & Paliwal, K., K. . (1996). *Automatic speech and speaker recognition - Advanced topics*. Boston: Kluwer academic publishers.

Li, K. P., & Porter, J. E. (1988, 11-14 Apr). *Normalizations and selection of speech segments for speaker recognition scoring.* Paper presented at the International Conference on Acoustics, Speech, and Signal Processing.

Luengo, I., Navas, E., Hernaez, I., Sanchez, J., Saratxaga, I., & Sainz, I. (2006). *Effectiveness of short-term prosodic features for speaker verification.* Paper presented at the IV Jornadas en tecnologia del Habla.

Luger, G. F. (2002). *Artificial intelligence: structures and strategies for complex problem solving* (4th ed.). London: Addison-Wesley.

Martin, A., & Przybocki, M. (2000). The NIST 1999 speaker recognition evaluation - overview. *Speech Communications, 31*, 225-254.

Martin, T. (2006). *Towards improved speech recognition for resource poor languages.* QUT, Queensland.

Martin, T., Baker, B., Wong, E., & Sridharan, S. (2006). A syllable-scale framework for language identification. *Computer Speech and Language, 20*, 125-127.

Martin, T., & Sridharan, S. (2005). *Cross-language acoustic model refinement for the Indonesian language.* Paper presented at the Proceedings of the International Conference on Acoustics, Speech, and Signal Processing.

Martin, T., Svendsen, T., & Sridharan, S. (2003). *Cross-lingual pronunciation modelling for Indonesian speech recognition.* Paper presented at the Conf. Eurospeech 2003, Geneva.

Martin, T., Wong, E., Baker, B., & Mason, M. (2004). *Pitch and energy trajectory modelling in a syllable length temporal frameowrk for language identification.* Paper presented at the ODYS-2004.

Martin, T., Wong, E., & Sridharan, S. (2006). *Towards improved assessment of phonotactic information for automatic language identification.* Paper presented at the IEEE 2006 Odyssey: The Speaker and Language Recognition Workshop.

Mason, M., Vogt, R., Baker, B., & Sridharan, S. (2004). *The QUT NIST 2004 speaker verification system: a fused acoustic and high-level approach.* Paper presented at the Tenth Australian International Conference on Speech Science and Technology, Macquarie University, Sydney.

Matejka, P., Cernocky, J., & Sigmund, M. (2004). *Introduction to automatic language identification.* Paper presented at the Conference Proceedings of Radioelektronika.

Millar, J. B., Vonwiller, J. P., Harrington, J. M., & Dermody, P. J. (1994). *The Australian national database of spoken language.* Paper presented at the Proceedings of the International Conference on Acoustics, Speech, and Signal Processing.

Moore, R. (2003). *A comparison of the data requirements of automatic speech recognition systems and human listeners.* Paper presented at the Eurospeech'03, Geneva.

Morgan, N., & Bourlard, H. (1995). Continuous speech recognition. *Signal Processing Magazine, IEEE, 12*(3), 24-42.

Muthusamy, Y. K., Barnard, E., & Cole, R. A. (1994). Reviewing automatic language identification. *Signal Processing Magazine, IEEE, 11*(4), 33-41.

Neagoe, V., & Ropot, A. (2004). *Concurrent self-organizing maps -a powerful artificial neural tool for biometric technology.* Paper presented at the Proceedings of World Automation Congress WAC'04, Seville.

Nealand, J. H., Bradley, A. B., & Lech, M. (2002). *Discriminative feature extraction applied to speaker identification.* Paper presented at the ICSP'02.

Ortega-Garcia, J., Gonzalez-Rodriguez, J., Marrero-Aguiar, V., Diaz-Gomez, J. J., Garcia-Jimenez, R., Lucena-Molina, J., et al. (1998). *AHUMADA: a large speech corpus in Spanish for speaker identification and verification.* Paper presented at the Proc. of the 1998 IEEE International Conference.

Paliwal, K. K. (1998). *Spectral subband centroid features for speech recognition.* Paper presented at the Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing.

Pelecanos, J., & Sridharan, S. (2001a, June 18-22, 2001). *Feature warping for robust speaker verification.* Paper presented at the Proceedings of a Speaker Odyssey, Crete, Greece.

Pelecanos, J., & Sridharan, S. (2001b). Rapid channel compensation for one and two speaker detection in the NIST 2000 speaker recognition evaluation. *Acoustics Australia, 29*(1), 17-20.

Price, R. C., Willmore, J. P., Roberts, W. J. J., & Zyga, K. J. (2000). *Genetically optimised feedforward neural networks for speaker identification.* Paper presented at the Fourth International Conference on Knowledge-based Intelligent Engineering Systems and Allied Technologies.

Rabiner, L., & Juang, B. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine 3*(1), 4-16.

Rabiner, L. R., & Huang, B. H. (1993). *Fundamentals of speech recognition*. Englewood cliffs, New Jersey: Prentice Hall.

Reynolds, D. (1997). *Comparison of background normalisation methods for text-independent speaker verification.* Paper presented at the European Conference on Speech Communication and Technology.

Reynolds, D., Quatieri, T., & Dunn, R. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing, 10*, 19-41.

Rohlicek, J. R. (1995). Modern methods of speech processing, chapter Word Spotting. In (pp. 136-140): Kluwer Academic publishers.

Rongqing, H., & Hansen, J. H. L. (2005). *Dialect/Accent Classification via Boosted Word Modeling*. Paper presented at the Proceedings of the International Conference on acoustics, speech, and signal processing.

Rose, P. (2006). *The Intrinsic forensic discriminatory power of diphthongs.* Paper presented at the Eleventh Australasian International Conference on Speech Science and Technology Auckland, New Zealand.

Rose, P., Kinoshita, Y., & Alderman, T. (2006). *Realistic extrinsic forensic speaker recognition with the diphthong /ai/.* Paper presented at the Eleventh Australasian International Conference on Speech Science and Technology Auckland, New Zealand.

Rose, P., Osanai, T., & Kinoshita, Y. (2003). Strength of forensic speaker identification evidence: multispeaker formant- and cepstrum-based segmental discrimination with a Bayesian likelihood ratio as threshold. *Speech Language and the Law, 10*(2), 179-202.

Rose, R. C., & Paul, D. B. (1990). *A hidden Markov model based keyword recognition system.* Paper presented at the International Conference on Acoustics, Speech, and Signal Processing.

Sawai, H. (1991). *Frequency-time-shift-invariant time-delay neural networks for robust continuous speech recognition.* Paper presented at the Proceedings of the International Conference on Acoustics, Speech, and Signal Processing.

Schultz, T., & Waibel, A. (2000). *Polyphone decision tree specialization for language adaptation.* Paper presented at the ICASSP 2000.

Sehgal, M. S. B., Gondal, I., & Dooley, L. (2004). *A hybrid neural network based speech recognition system for pervasive environments.* Paper presented at the INMIC 2004.

Shawe-Taylor, J., & Cristianini, N. (2000). *Support vector machines and other kernel-based learning methods*: Cambridge University Press.

Tan, B. T., Minyue, F., Spray, A., & Dermody, P. (1996). *The use of wavelet transforms in phoneme recognition.* Paper presented at the Fourth International Conference on Spoken Language Processing, Philadelphia, PA, USA.

Tanabian, M. M., & Goubran, R. A. (2005). *Speech accent identification with vocal tract variation trajectory tracking using neural networks.* Paper presented at the Proceedings of the 2005 IEEE International Conference on Computational Intelligence for Homeland Security and Personal Safety.

Tey, W., Jong, N., & Togneri, R. (1996). *Investigation of speech and speaker recognition based on trajectory modelling of utterances.* Paper presented at the Sixth Australian International Conference on Speech Science and Technology, Adelaide, Australia.

Thambiratnam, K., Martin, T., & Sridharan, S. (2004). *A study on the effects of limited training data for English, Spanish and Indonesian keyword spotting.* Paper presented at the 8th European Conference on Speech Communication and Technology (Eurospeech), Geneva, Switzerland.

Thambiratnam, K., & Sridharan, S. (2005). *Dynamic match phone-lattice searches for very fast and accurate unrestricted vocabulary keyword spotting.* Paper presented at the Proceedings of the International Conference on Acoustics, Speech, and Signal Processing.

Togneri, R., & Deng, L. (2004). *Use of neural network mapping and extended Kalman filter to recover vocal tract resonances from the MFCC parameters of speech.* Paper presented at the Proceedings of International Conference on Speech and Language Processing.

Togneri, R., & Li, D. (2001). *An EKF-based algorithm for learning statistical hidden dynamic model parameters for phonetic recognition.* Paper presented at the Proceedings of the International Conference on Acoustics, Speech, and Signal Processing.

Toledano, D. T., Gomez, L. A. H., & Grande, L. V. (2003). Automatic phonetic segmentation. *Speech and Audio Processing, IEEE Transactions on, 11*(6), 617-625.

Too, C., Chao, H., Chang, E., & Jingehan, W. (2001). *Automatic accent identification using Gaussian mixture models.* Paper presented at the ASRU'01.

Torres-Carrasquillo, P., Singer, E., Kohler, M., Greene, R., Reynolds, D., & Deller, J. (2002). *Approaches to language identification using gaussian mixture models and shifted delta cepstral features.* Paper presented at the Proceedings of the International Conference on Spoken Language Processing, Denver.

Tranter, S. E., & Reynolds, D. (2006). An overview of automatic speaker diarisation systems. *IEEE Transactions on Speech and Audio Processing, Special Issue on Rich Transcription, Vol 14*(5), 1557-1565.

Trentin, E., & Gori, M. (2001). A survey of hybrid ANN/HMM models for automatic speech recognition *Neurocomputing, 37*(1), 91-126.

Trentin, E., & Gori, M. (2003). Robust combination of neural networks and hidden Markov models for speech recognition. *Neural Networks, IEEE Transactions on, 14*(6), 1519-1531.

Wegmann, S., McAllaster, D., Orloff, J., & Peskin, B. (1996). *Speaker normalization on conversational telephone speech.* Paper presented at the International Conference on Acoustics, Speech, and Signal Processing.

Weibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. J. (1989). Phoneme recognition using time delay neural networks. *IEEE Trans. Acoustics, Speech Signal Processing, 37*, 328-339.

Wong, E., & Sridharan, S. (2001). *Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification.* Paper presented at the Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing.

Wong, E., & Sridharan, S. (2002a). *Methods to improve Gaussian mixture model based language identification system.* Paper presented at the International Conference on Spoken Language Processing.

Wong, E., & Sridharan, S. (2002b). *Utilise vocal tract length normalization for robust automatic language identification.* Paper presented at the Australian International Conference on Speech Science and Technology.

Wong, E., & Sridharan, S. (2003). *Three approaches to multilingual phone recognition.* Paper presented at the Proceedings of the International Conference on Acoustics, Speech, and Signal Processing.

Young, S. (1996). A review of large-vocabulary continuous-speech. *Signal Processing Magazine, IEEE, 13*(5), 45.

Young, S. J., Brown, M. G., Foote, J. T., Jones, G. J. F., & Jones, K. S. (1997). *Acoustic indexing for multimedia retrieval and browsing.* Paper presented at the Proceedings of the International Conference on Acoustics, Speech, and Signal Processing.

Zhou, J., Liu, J., Song, Y., & Yu, T. (1998). *Keyword spotting based on recurrent neural network.* Paper presented at the Fourth international Conference on Signal Processing

Zissman, M. A. (1995). *Language identification using phone recognition and phonotactic language modelling.* Paper presented at the International Conference on Acoustics, Speech, and Signal Processing.

Zissman, M. A. (1996). Comparison of four approaches to automatic language identification of telephone speech. *Speech and Audio Processing, IEEE Transactions on, 4*(1), 31-44.

Zissman, M. A., & Singer, E. (1994). *Automatic language identification of telephone speech messages using phoneme recognition and N-gram modeling.* Paper presented at the Proceedings of the International Conference on Acoustics, Speech, and Signal Processing.

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION
DOCUMENT CONTROL DATA

1. PRIVACY MARKING/CAVEAT (OF DOCUMENT)

2. TITLE

A Review of Contributions by Australian Research Institutions into Speech Processing

3. SECURITY CLASSIFICATION (FOR UNCLASSIFIED REPORTS THAT ARE LIMITED RELEASE USE (L) NEXT TO DOCUMENT CLASSIFICATION)

Document	(U)
Title	(U)
Abstract	(U)

4. AUTHOR(S)

Trevor Chi-Yuen Tao

5. CORPORATE AUTHOR

DSTO Defence Science and Technology Organisation
PO Box 1500
Edinburgh South Australia 5111 Australia

6a. DSTO NUMBER	6b. AR NUMBER	6c. TYPE OF REPORT	7. DOCUMENT DATE
DSTO-TN-0837	AR-014-256	Technical Note	August 2008

8. FILE NUMBER	9. TASK NUMBER	10. TASK SPONSOR	11. NO. OF PAGES	12. NO. OF REFERENCES
2008/1017774	07/020	DSD	42	107

13. URL on the World Wide Web

http://www.dsto.defence.gov.au/corporate/reports/DSTO-TN-0837.pdf

14. RELEASE AUTHORITY

Chief, Command, Control, Communications and Intelligence Division

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT

Approved for public release

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, EDINBURGH, SA 5111

16. DELIBERATE ANNOUNCEMENT

No Limitations

17. CITATION IN OTHER DOCUMENTS Yes

18. DSTO RESEARCH LIBRARY THESAURUS http://web-vic.dsto.defence.gov.au/workareas/library/resources/dsto_thesaurus.shtml

Select 3-5 descriptors/keywords from the library thesaurus

19. ABSTRACT

This report is a survey of contributions by various research institutions within Australia into several important applications of speech processing, such as speech and speaker recognition. The purpose of this report is to give a rough snapshot of where a number of individual research institutions stand. For each application, a number of research papers within Australia are discussed in detail. Although much of the above research is directed towards simple tasks there are a number of significant contributions from various Australian research institutions. Some systems, particularly those from QUT, have achieved state-of-the-art performance.